DOCUMENT RESUME

ED 317 580 TM 014 629

AUTHOR	Richards, Llyn, Ed.; Croft, Cedric, Ed.
TITLE	The Best of "Set" Assessment.
INSTITUTION	Australian Council for Educational Research, Hawthorn.; New Zealand Council for Educational Research, Wellington.
PUB DATE	89
NOTE	108p.
AVAILABLE FROM	Set, New Zealand Council for Educational Research, Box 3237, Wellington, New Zealand. Set, Australian Council for Educational Research, Box 210, Hawthorne, Victoria 3122, Australia.
PUB TYPE	Collected Works - General (020) -- Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Achievement Tests; Anthologies; Criterion Referenced Tests; *Educational Assessment; Elementary Secondary Education; *Evaluation Methods; Foreign Countries; Intelligence Tests; Item Banks; *Measurement Techniques; Nonverbal Tests; Observation; Scores; *Standardized Tests; Test Bias; *Test Use; Writing Evaluation
IDENTIFIERS	Australia; New Zealand

ABSTRACT

This package contains articles in three general areas: items covering measurement topics; brief and practical guides on measurement techniques; and professional reading on broader assessment issues. The purpose of these publications by "Set" is to provide research information to teachers. The initial article, "Overview of Issues in School Assessment" (B. McGaw), was written for this compilation. The other 13 items were all published in "Set" between 1978 and 1987. They are: (1) "Achievement Test Scores in Perspective" (W. Turnbull); (2) "The Foundations of School Testing" (C. Croft); (3) "Test Evaluation Sheet" (S. Larsen and D. Hammill); (4) "Assessing What They've Learned" (W. B. Elley); (5) "Criterion-Referenced Measurement" (G. Rowley and C. Macpherson); (6) "Investing in Item Banks" (N. Reid); (7) "Combining Scores" (A. Gilmore); (8) "Evaluating Writing" (D. Phillips); (9) "Observation: The Basic Techniques" (B. McMillan and A. Meade); (10) "One Extreme to the Other: A Report on Profile Reports" (G. Withers); (11) "Non-Verbal Tests in Schools" (C. Croft); (12) "Does Intelligence Equal Learning Ability?" (J. Jenkinson); and (13) "Test Bias! Test Bias!" (N. Reid and A. Gilmore). (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED317580

THE BEST OF set ASSESSMENT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

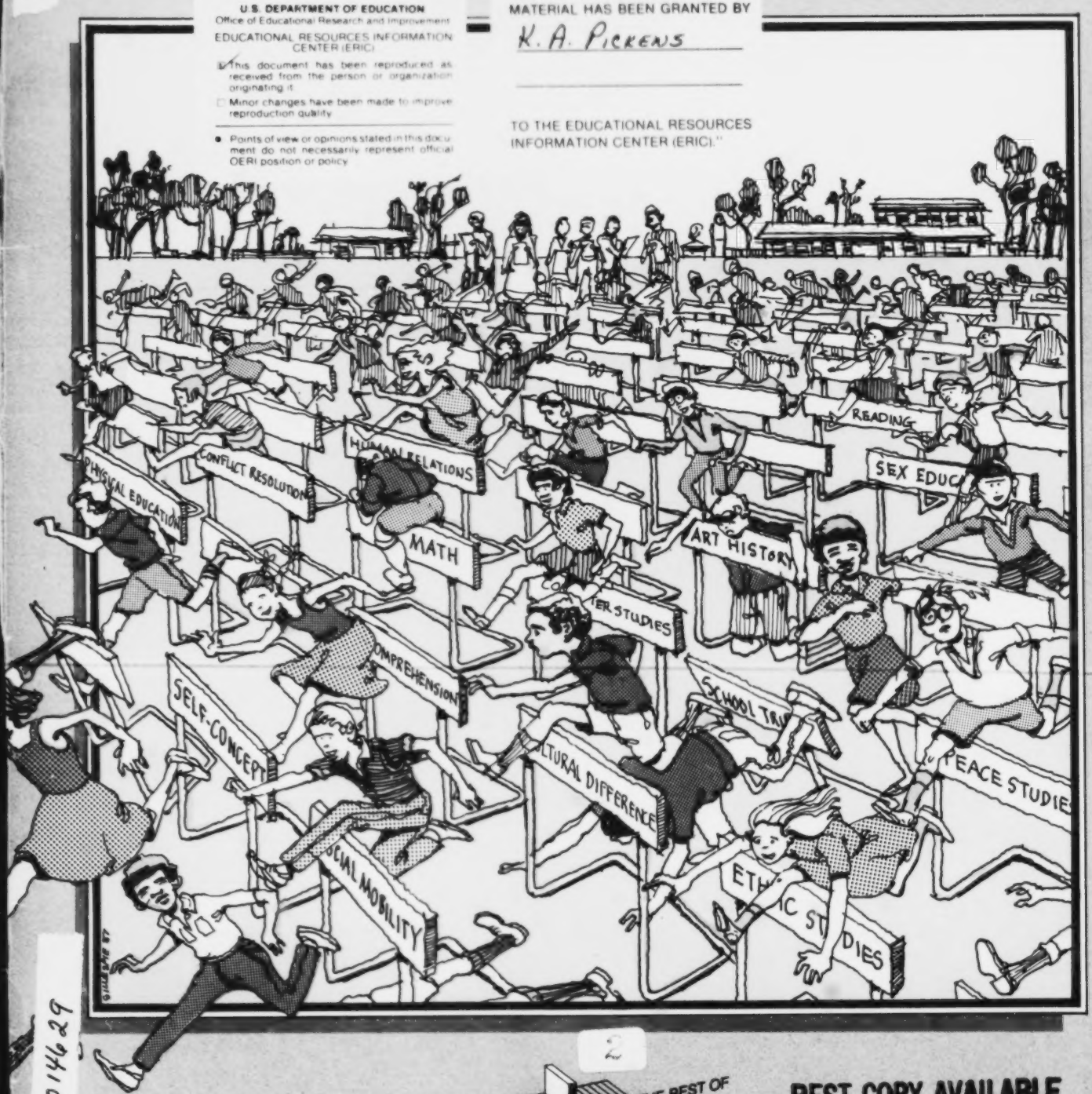
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. A. Pickens

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



TM014629

THE BEST OF set ASSESSMENT

BEST COPY AVAILABLE

set: *research information for teachers*, is published twice a year by the New Zealand Council for Educational Research and the Australian Council for Educational Research.

General Editor: Llyn Richards
Special Editor: Cedric Croft
Cover Design: John Gillespie
Australian Editor: Peter Jeffery

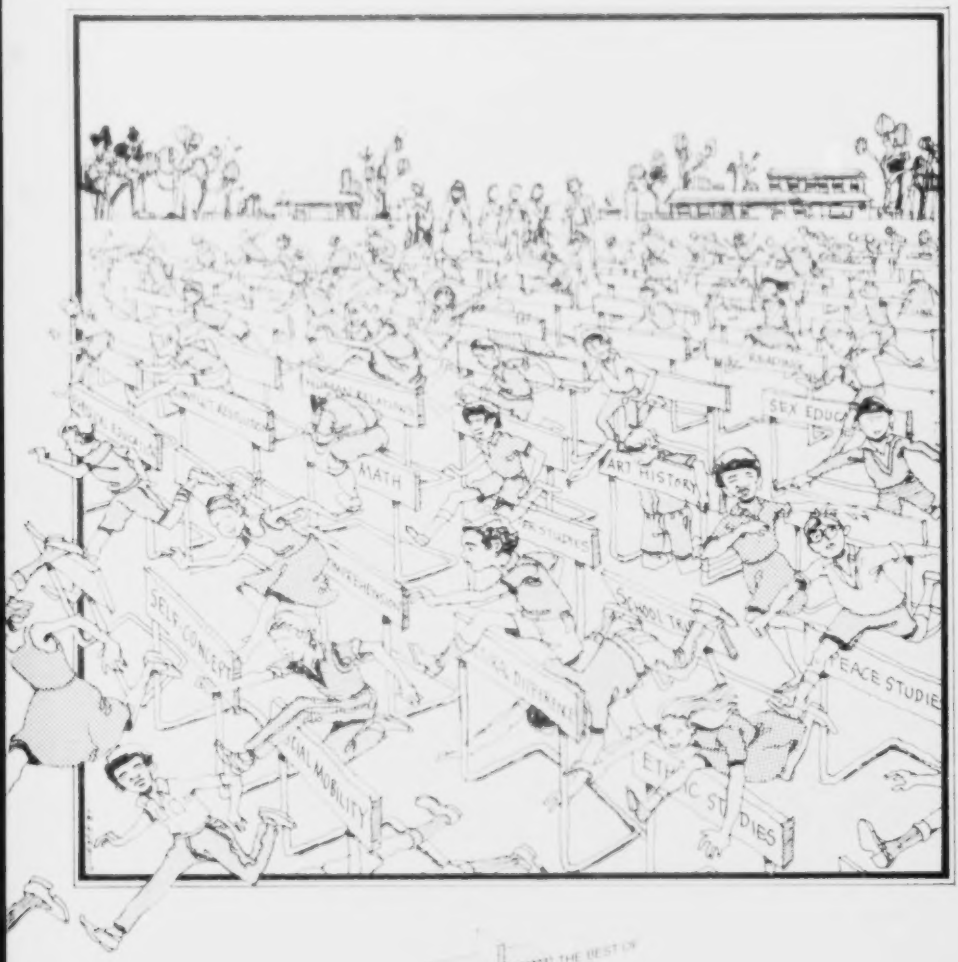
Copyright: NZCER/ACER 1987
Copying by arrangement:
Achievement Scores in Perspective by Bill Turnbull (Copyright held by Educational Testing Services, Princeton, New Jersey, USA).
Copying Permitted: Copyright on all other items is held by NZCER and ACER who grant to all people actively engaged in education the right to copy them in the interests of better teaching; simply acknowledge the source.

Subscription orders, and inquiries, to:

New Zealand
set,
NZCER
Box 3237
Wellington

Australia
set
ACER
Box 210
Hawthorn
Victoria, 3122.

THE BEST OF set ASSESSMENT



In this package are gathered together

- **items covering measurement topics;**
- **brief and practical guides on measurement techniques;**
- **professional reading on broader assessment issues.**

We begin with *Overview of Issues in School Assessment*, an item specially written for this *Best of set*. Then come thirteen items, all published previously in *set* between 1980 and 1987, and reprinted without alteration. Some still have their old *set* item numbers on them but they are arranged in this package under four headings:

Assessment Issues and Measurement Concepts

Overview of Issues in School Assessment

Achievement Test Scores in Perspective

Foundations of School Testing

Test Evaluation Sheet

Measurement and Assessment Techniques

Assessing What They've Learned

Criterion Referenced Tests

Investing in Item Banks

Combining Scores

Evaluating Writing

Observation: The Basic Techniques

Reporting

One Extreme to the Other: A Report on Profile Reports

Assessment, Abilities and Culture

Non-Verbal Tests in Schools

Does Intelligence Equal Learning Ability?

Test Bias! Test Bias!

Summaries

Assessment Issues and Measurement Concepts

Overview of Issues in School Assessment.
Barry McGaw.

A discussion of the *purposes*, the *qualities*, the *forms* and the *content* of assessment. Various approaches to aggregating or summarising results are considered, and professional obligations in assessment outlined. The fundamentals of item response theory are covered as this technique offers an integration of criterion-referenced and norm-referenced assessment.

Achievement Test Scores in Perspective.
Bill Turnbull.

The focus is on three virtues of standardised tests: accuracy, objectivity, and comparability; and three fallacies that have helped promote the misuse of test scores: the micrometer fallacy, the whole person fallacy and the equal preparation fallacy. Legitimate uses of standardised tests are outlined and cautions raised on inappropriate uses.

Foundations of School Testing.
Cedric Croft.

This item outlines and illustrates key elements of the concepts of *validity*: content, concurrent, predictive, construct; *reliability*: test-retest, alternate forms, split-half, Kuder-Richardson; and *usability*. Illustrations are built around common standardised tests and likely examining and testing situations in schools. Much of the discussion underpins reference to these concepts in other items.

Test Evaluation Sheet.
Cedric Croft.

Designed to accompany the previous article, this sheet provides a convenient basis for evaluating the qualities of published tests and possibly some teacher-made tests too. A worked example showing an entry for a published test is included.

Measurement and Assessment Techniques

Assessing What They've Learned.
Warwick Elley.

A brief guide to the essential aspects of planning and writing a good test or examination. Examples of four types of objective items are included, with additional comment on the advantages claimed for multiple-choice items. Guidelines are given for preparing essay questions. The length, structure and organisation of test and examination papers is also covered.

Criterion Referenced Tests.
Glenn Rowley and Colin MacPherson.

This item sets out the major characteristics, strengths and weaknesses of criterion referenced tests. Four fundamental requirements of these tests are outlined and 'the standard setting problem' is discussed. Criterion referenced tests and norm referenced tests ask different questions, but are complementary, each may illuminate different aspects of a child's development.

Investing in Item Banks.
Neil Reid.

Item banks have an instant appeal to busy educators who must juggle the continual demands of teaching and assessing. This article describes their characteristics, follows their development and considers the vitally important questions of classification and retrieval. The actual operation and future of item banks are discussed, and the few available collections of items in Australia and New Zealand listed.

Combining Scores.
Alison Gilmore.

There will be occasions when many educators will need to combine scores from several sources. How valid are these composite scores? Which types of scores may be combined legitimately? What factors should be considered when scores are to be combined? Illustrations are included of common situations where scores must be combined. Worked examples show the actual process.

Evaluating Writing.
David Philips

The *why* and *how* of assessing writing continue to be difficult questions. After considering some basic aspects of writing this overview discusses methods of marking, and outlines features of analytic and holistic schemes. Examples of analytic scales and a checklist for writing assessments are included.

Observation: The Basic Techniques.
Bruce McMillan and Anne Meade

The uses of observation and description of eight techniques. Most of the discussion and examples are drawn from early childhood education but the techniques described are applicable to any situation requiring accurate observing of behaviour. For internal assessment in secondary schools and for a broader range of assessments, the techniques described here could be applied in the laboratory, in practical subjects or in field-work.

Reporting

One Extreme to the Other: A Report on Profile Reports.
Graeme Withers

Profile reports appear to offer schools the means of giving information about a wide range of achievement and activity. Essential features are illustrated, seven profile types are identified and a number of forms shown. Eight important features are outlined and comments made on the nature, content and limitations of profiles.

Assessment, Abilities and Culture

Non-Verbal Tests in Schools.
Cedric Croft

This article examines three common assumptions: (1) non-verbal tests are measures of general intelligence; (2) non-verbal tests are more valid measures of the school potential of low achievers than are verbal tests; (3) non-verbal tests are culture free. Characteristics of non-verbal, non-language and performance tests are considered. Legitimate uses of non-verbal tests within schools are discussed.

Does Intelligence Equal Learning Ability?
Jo Jenkinson

Educators may assume a perfect relationship exists between intelligence and learning ability, but the relationship is stronger when understanding of principles is required, when learning is moderately complex relevant to a given age, and when learning is in its early stages. Traditional measures of intelligence are not designed to predict learning styles, to measure processes, nor to assess the particular abilities needed to learn new tasks. They predict outcome rather than process.

Test Bias! Test Bias!
Neil Reid and Alison Gilmore

Five major classes of potential test bias are identified: content, language, test atmosphere, prediction and selection, social consequences; the nature and likely outcomes of each are explored; guidance is given on how to eliminate content bias. But discussions of bias are likely to be political in nature rather than academic. Many factors leading to test bias probably have yet to be identified, let alone controlled or eliminated.

Overview of Issues in School Assessment

by Barry McGaw

Australian Council for Educational Research

Purposes of Assessment

ASSESSMENT IN EDUCATION serves several purposes. It can provide diagnoses of learning and teaching deficiencies. It can establish the level of achievement of a student or a group of students with reference to defined standards or with reference to the achievements of other students. It can indicate the areas in which learning and teaching are in need of improvement. It can provide the basis of an overall judgment of the quality of learning and teaching.

Diagnosis

In many ways, the most important purpose is diagnosis. Only by careful diagnostic assessment can a teacher identify the particular misconceptions a student has developed and then seek to remove them. The task of diagnosis is more straightforward in highly structured subjects such as mathematics but it is possible in all, so long as the teacher is clear about what knowledge and skills are to be developed.

Systematic use of diagnostic assessment can identify for the teacher deficiencies in teaching. If a particular misconception is common to a reasonable number of students then a teaching deficiency and not just a learning deficiency is indicated. This can alert the teacher to the need for general supplementary instruction and to the need for a different instructional pattern with future groups of students.

Determining Levels of Achievement

The second broad purpose of assessment is to determine the levels of achievement of individuals or groups. That can be done in either of two ways. One approach is usually called 'criterion-referenced' assessment because it involves the definition of standards of achievement as the criterion against which performances are judged. The other approach is usually called 'norm-referenced' assessment because it involves comparison of performances with the average or norm achievement of some reference group of students.

These approaches appear to be fundamentally different, one using some *absolute* set of standards as the point of reference, the other using actual achievements of other students for *comparison*. In fact, there is some overlap. In defining criterion levels of achievement, say for mathematics in Year 5 (Standard 4) in the primary school, the choice must be based on some consideration of what it is reasonable to expect of students of this age and experience and thus on some consideration of what they typically achieve. To this extent, the criteria are normatively defined.

For norm-referenced assessment, the choice of reference group is important. The most restricted case involves the comparison of a student's achievement only with that of others in the same class. If the achievements are established on a particular test prepared by the class teacher and administered only to that class then, of course, no more is possible. The result, however, may be quite misleading if the class is atypical in any way. For example, if a student is declared to be 'below average' or of low rank in a class that is as a whole well above average, it would be wrong to conclude that the student is in any general way 'below average'. Without evidence of the high standing of the whole class, the student, the student's parents and even the teacher may not be able to avoid this incorrect interpretation of the normative assessment of the student's performance.

Evidence about the general level of the student's and the class's performances can be obtained with standardised tests. With such tests, the levels of performances of a standard group of students, genuinely representative of the relevant population, have been determined. The individual can then be compared to the population and not just to the local class. One important disadvantage of this approach is that the standardised test might not match the teacher's instructional purposes as well as the teacher's own test could. A mix of test types is, therefore, of most value.

Both norm-comparisons among students and criterion-comparisons with defined standards can be achieved simultaneously. The purposes may be different but the strategies are not as different as many have supposed. Norm-referenced tests are designed to spread students out to establish the nature of differences among them. When this is done it establishes also the types of things that only the best achievers can do, those that the average students can do, and so on. The result can be an ordered set of criteria. The technical basis for achieving such an integration of norm-referenced and criterion-referenced assessment is outlined later.

Judging the Quality of Learning and Teaching

There is another distinction in purpose that cuts across the one already made. This distinction was first introduced in discussions of curriculum evaluation. It is the distinction between formative and summative assessment. The purpose of *formative* assessment is to improve teaching and learning. The purpose of *summative* assessment is to certify levels of achievement in some final judgment on a student's progress at a particular stage of education such as in an end-of-year report or an end-of-schooling certificate or in some final judgment on the teacher's performance.

Diagnostic, criterion-referenced and norm-referenced assessments can all serve both formative and summative purposes. For formative purposes, however, it must be said that diagnostic and criterion-referenced assessment are more relevant. They identify more clearly the deficiencies that need to be dealt with.

Necessary Properties of Instruments

ALL ASSESSMENT INSTRUMENTS must satisfy certain technical requirements if they are to achieve their purposes. The traditional expression of these requirements was to say that instruments must be valid and reliable. The validity criterion remains much as it has been for several decades. The reliability criterion has been recast in important ways over the last decade or so through some significant developments in psychometric theory.

Validity

Questions of validity are about the appropriateness of inferences that may be based on test scores. Three different conceptions of validity have been defined to clarify the notion.

Content validity depends on the adequacy with which the performances demanded in the test are representative of the knowledge and skills in the domain to be assessed. This requires that the domain be defined, that the method of selecting the tasks to be used in assessment be clear, and that the representativeness of the chosen tasks be established. This can involve setting out some map of the curriculum and a corresponding map of the test's content to ensure that it corresponds with the curriculum emphasis in the intended ways. In practice, many teachers and others concern themselves only with the face validity of an instrument, asking only whether it looks to be assessing the appropriate things. A more systematic analysis of content is necessary to be confident.

Criterion-related validity depends on the adequacy with which test scores can be used to make inferences about an individual's probable standing on some other variable, the criterion. If the criterion is some future level of performance, such as results in the first year of higher education, and the test is of current performance, such as Year 12 (Form 7) assessments, the correlation between the two sets of scores is a measure of the criterion-related validity of the current assessments. In this case, it can be described as *predictive* validity. In other cases the concern may be with finding an efficient way of assessing current status without direct and elaborate assessment. For example, a test may be developed as a quick way of identifying the extent and nature of a student's reading difficulties. In this case, the correlation of scores on the test with other more extensive assessments of reading performance is a measure of criterion-related validity, here called *concurrent* validity.

Construct validity depends on the extent to which an instrument measures the theoretical construct it is intended to measure. A theoretical construct is an idea used to explain or organise knowledge. Terms such as 'mathematical aptitude', 'anxiety' and 'reading readiness' are labels for such constructs. Evidence of construct validity cannot be provided by a single correlation with some other measure or by systematic analysis of the content of a test. It depends on an accumulating body of research on the topic.

Most of the debate about school assessment is about validity. Those who criticise attempts to monitor the performance of schools or educational systems with system-wide testing programmes, for example, are usually concerned that the focus of the testing is too narrow and fails to reflect the full range of objectives to which the educational effort is directed. Those who criticise external examinations are usually concerned about failures of the examinations to test an adequate sample of the performances required by the curriculum and their propensity to test other less relevant things such as capacity to work quickly under pressure.

Precision

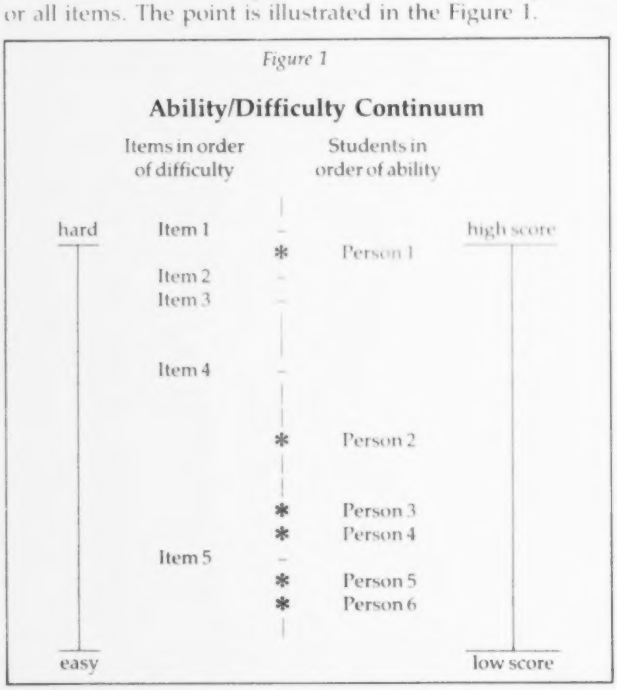
Traditional discussions of the precision with which an instrument measures have been cast in terms of reliability. The correlation between scores obtained by a group of students measured twice with the same test is defined to be the test-retest reliability of the test. The correlation between scores of a group of students measured with two forms of the same test is defined as the equivalent-forms reliability. The correlations between scores on all pairs of items within a test are accumulated into a single index of internal consistency which is the most commonly used index of reliability. It provides an index of how well the components of a test measure the same thing.

All of these indices of reliability are indirect measures of the measurement error with which assessments of individual performances are made. They are derived from a simple mathematical model of a test score which asserts that the observed score of a student reflects the student's true score, and an error of measurement, which may be positive or negative. One problem is that the model assumes that all individuals are measured with the same precision, regardless of how high or low their score is.

Contemporary test theory offers us another way to think about the basis of scores achieved on a test. Instead of concentrating on total test score and taking it to be the sum of a true score and an error score, it uses a model of what happens when an individual answers a single item. For this reason it is called 'item-response test theory'. It takes individuals and items to be on a single dimension, the individuals according to their ability and the items according to their difficulty. If an item's difficulty is right at the level of a person's ability, the person is taken to have a probability of 0.5 (that is a 50:50 chance) of answering the item correctly. If the item is above the person's ability, the person is taken to have a probability of less than 0.5 of being correct, with the probability declining the further the item is above the person's ability. For items below the person's ability, the person is taken to have a probability of greater than 0.5 of being correct, with the probability rising the further the item is below the person's ability.

We need not consider the mathematical model employed to represent this interaction between persons and items. The model expresses the probability of a person answering correctly as a function of the person's ability and the item's difficulty. The only point to note is that by considering the actual pattern of correct and incorrect responses of a group of individuals to a set of items, it is possible to estimate the ability of each person and the difficulty of each item. Furthermore, it is possible to estimate the standard error with which each estimate of ability and difficulty is made. There is no need to assume that the precision is the same for all persons or all items. The point is illustrated in the Figure 1.

Figure 1



With this approach to the estimation of persons' abilities and items' difficulties, it is possible to obtain a clear definition of the ability/difficulty continuum. The ordering of the items by difficulty can help a teacher understand the relative

levels of different tasks and can help with the definition of criteria for criterion-referenced assessment. The location of students by ability on the same scale provides, strange though it may seem, a direct criterion-referenced assessment of their performances. At the same time, the measurement allows for comparison among students and, if the levels of performance of reference groups have been established, it allows also for normative assessment with respect to those reference groups. Instead of forcing a choice between criterion-referenced and norm-referenced assessment, this approach allows either or both to be undertaken according to the teacher's or the student's requirements at a particular time.

An increasing number of standardised tests now provide information about test properties and about individual's test scores in terms of this approach. Classical test theory with its notions of reliability and common errors of measurement offers only normative assessment. Item-response test theory offers an integration of norm-referenced and criterion-referenced assessment.

Forms of Assessment

Informal Observations

TEACHERS MAKE INFORMAL ASSESSMENTS OF students' achievements all the time. Every time a teacher asks a question, the answers are evaluated and a decision taken about whether to probe deeper to refine the assessment, to accept the response as evidence of understanding and so to continue the instruction, or to accept it as evidence of lack of understanding and so to repeat or otherwise reinforce the original instruction.

This type of assessment is an essential element of all good teaching and should not be undervalued simply because it is not structured and formal. To use it well requires considerable professional skill. To use it wisely requires that it not be the only form of assessment used.

Teacher-made Tests

Teachers can find it difficult to accumulate an overall assessment of progress from the questions and answers that occur in the ebb and flow of instruction. More formal and structured assessments can provide the overall view. To achieve this teachers design and administer 'tests'. The test might involve completion of practical work in a laboratory, solution of problems or writing of an essay in class. It might involve similar work completed at home or under examination conditions. Some tests will be objective in the sense that it is clear what constitutes a correct answer. Others will be subjective in the sense that the teacher must exercise judgment in determining the adequacy of the response. Neither form is adequate alone.

External Assessments

In some circumstances, assessments can be based on tasks not designed by the individual teacher. The teacher may choose to use a standardised test in order to compare achievements of students in her class with those of a reference group. The choice of standardised test should then be based on the relevance of its content and the appropriateness of the reference group on which the normative information has been established. The manual for the test should provide explicit information about the sample of persons whose performances provide the norms, make clear how old the data are, and give evidence of the reliability and the precision of the test.

In the case of public examinations, the design of the test is not in the hands of the teacher but fidelity to a publicly defined curriculum is at least intended. In Australia public examinations remain only at the end of secondary education where they serve in the selection of students for admission to higher education. In New Zealand there are now public examinations in the third and fifth year of secondary education, with an externally moderated certificate in the fourth year. In the United Kingdom a new common curriculum is being defined for primary and secondary education and public assessment procedures are being planned for various grade levels. In the USA, without such a curriculum definition, more general standardised tests are used. The US practices and the UK developments reflect an external imposition of assessments to provide public reports on the achievements of schools and the educational system as a whole. Pressure for this kind of assessment may well grow in Australia and New Zealand as well.

Content of Assessment

THE CONTENT OF ASSESSMENTS can obviously be described by the subject covered. There are tests of reading, of science, of mathematics and so on. A more general description of content can be provided using the type of outcome or process being assessed.

Cognitive Outcomes

The most common forms of assessment focus on outcomes and then only on cognitive outcomes. Teachers are usually clearest about the intellectual skills that they are seeking to develop, though they often build tests which concentrate on the recall of information only. Higher order cognitive skills are important in teaching and also in assessment. Tests should require more than recall of factual information. They must also require the use of skills such as making inferences, analysing, synthesising and evaluating.

Psychomotor Outcomes

In many cases, there are important psychomotor objectives which should be the subject of assessment as well as of teaching. Laboratory procedures in science, quality of technique in art, technology, woodwork and so on are all important and relevant criteria for judging performance and all reflect psychomotor objectives.

Affective Outcomes

Affective objectives can be more problematic. If 'liking for science' is a teacher's objective for students, should a student be judged negatively if that outcome is not achieved, regardless of the level of the student's achievement of other objectives? Is it indoctrination to demand that particular attitudes and dispositions be developed and displayed? Should cognitive and psychomotor performances be the only bases on which the student is judged? Is the legitimacy of judging a student clearer with an affective objective like 'tolerance for ethnic differences'? This is not the place to engage in a detailed debate. The point is raised to make clear that school assessment raises ethical as well as technical questions and to emphasise the problematic nature of assessment of affective objectives.

Learning Process

Recent developments in assessment in an alternative upper secondary programme in Victoria have seen the introduction of a new emphasis on the learning processes themselves as well as on the outcomes. The approach has been labelled 'goal-based' assessment because it involves an initial negotiation between teacher and student about the goals to be achieved in the particular course.

The goals are statements of content and expected performance and also statements of the activity designed to achieve the performance. In an English unit, for example, the goal will specify a workload (so many words or essays in a certain time), a standard (clear expository writing), and a process (developed through successive draftings, corrections, editing and rewritings). Satisfactory performance requires satisfaction of all aspects.

Summarising Assessments

FOR EACH INDIVIDUAL STUDENT, many assessments can be obtained within individual courses and across courses. How this information might be summarised is an important question. A traditional approach has been to produce some aggregate or average as a summarising statistic to which all the information might be reduced. Reducing the results to a single index like this, of course, loses much information. An alternative is to retain at least some separate measures and to present the information as a profile of results thus preserving details.

Aggregates

In some Australian states the most prominent use of a single aggregate as a summary of performance is in producing tertiary entrance scores. A student's results at the end of secondary education, in some number of subjects, are aggregated to produce a single score. From these an overall ranking of candidates for admission to higher education is produced. The reduction of the information to a single aggregate and a single order of merit is justified on the grounds of the competitiveness with which access to higher education is sought.

What is lost in the aggregate are the multidimensionsality of the original data. A student with high scores in humanities and low scores in mathematics and science can obtain the same average as a student with high scores in mathematics and science and low scores in humanities and as another student with average scores in all. Given the specialised nature of study in higher education, it makes no sense to consider these applicants equal. One is much better prepared to study science and another much better prepared to study the humanities. One of the more powerful arguments for ignoring these differences and persisting with a single aggregate is that the alternative would require additional constraints upon subject choices that students might make in upper secondary education. The use of an aggregate, without consideration of the mix of subjects from which it is derived, is seen as a substantial and important concession to freedom of choice in secondary education. The loss of information in the reduction of the assessments to a single aggregate is then accepted as an acceptable price to pay.

Profiles

There is a growing reaction against the use of a single summarising aggregate or average in favour of a full profile of results. Reporting a profile of performances is by no means a new practice. Reports to parents usually provide such information and give separate results for separate subjects. Within any particular subject it is common to reduce all performances to a single score, but even that is not necessary. Some primary schools, for example, report with a more fine grained analysis on components of subjects such as English, preserving information about reading, writing, speaking, listening and so on.

Even for admission to higher education, profiles have been used in the past. Admission to university in some countries was obtained through the achievement of 'matriculation', obtained with some minimum mix of results such as two Bs and three Cs. It was only when competition for places became more intense and the need arose for a more precise ranking, or at least an apparently more precise ranking, that a single order of merit was produced on the basis of a single aggregate for each student.

There are increasing demands for a return to the use of profiles as a way of retaining some of the richness of the data which is lost through reduction to an aggregate. Unfortunately, much of the argument for profiles says no more than: more is better. There is a risk that extensive information will swamp the user and result in unclear and irreproduceable decisions. Unless clear decision rules are made, there is no way of ensuring that the decision for an individual student will not be simply the idiosyncratic judgment of the person who happens to deal with the information. Where no selection is involved, the profile can more readily stand as an appropriate rich report of performances.

As an example of the type of decision rule which might be made for using a profile for selection purposes, consider a simple case: reduce to a two dimensional profile the humanities and mathematics/science scores; the two pieces of information for each student can then be used in different ways for different purposes. For admission to an engineering programme, some minimum score on the humanities scale may be required and, for all those who satisfy that minimum, a rank order on the mathematics/science scale might be established to determine admissions. For economics, where both verbal and quantitative skills are required, an average of the two results may be used. For admission to a humanities programme, results on the humanities scale alone might be used. With further pieces of information in the profile, the decision rules will become more complex but unless they are made explicit they will not be reproduceable. More will be better only if it is clear what can be done with it.

Professional Obligations

EVERYONE WHO TEACHES has a professional obligation to assess performance. It is necessary to monitor the effectiveness of one's teaching, as well as to inform learners, (and any others responsible for them such as parents,) of the success of their learning.

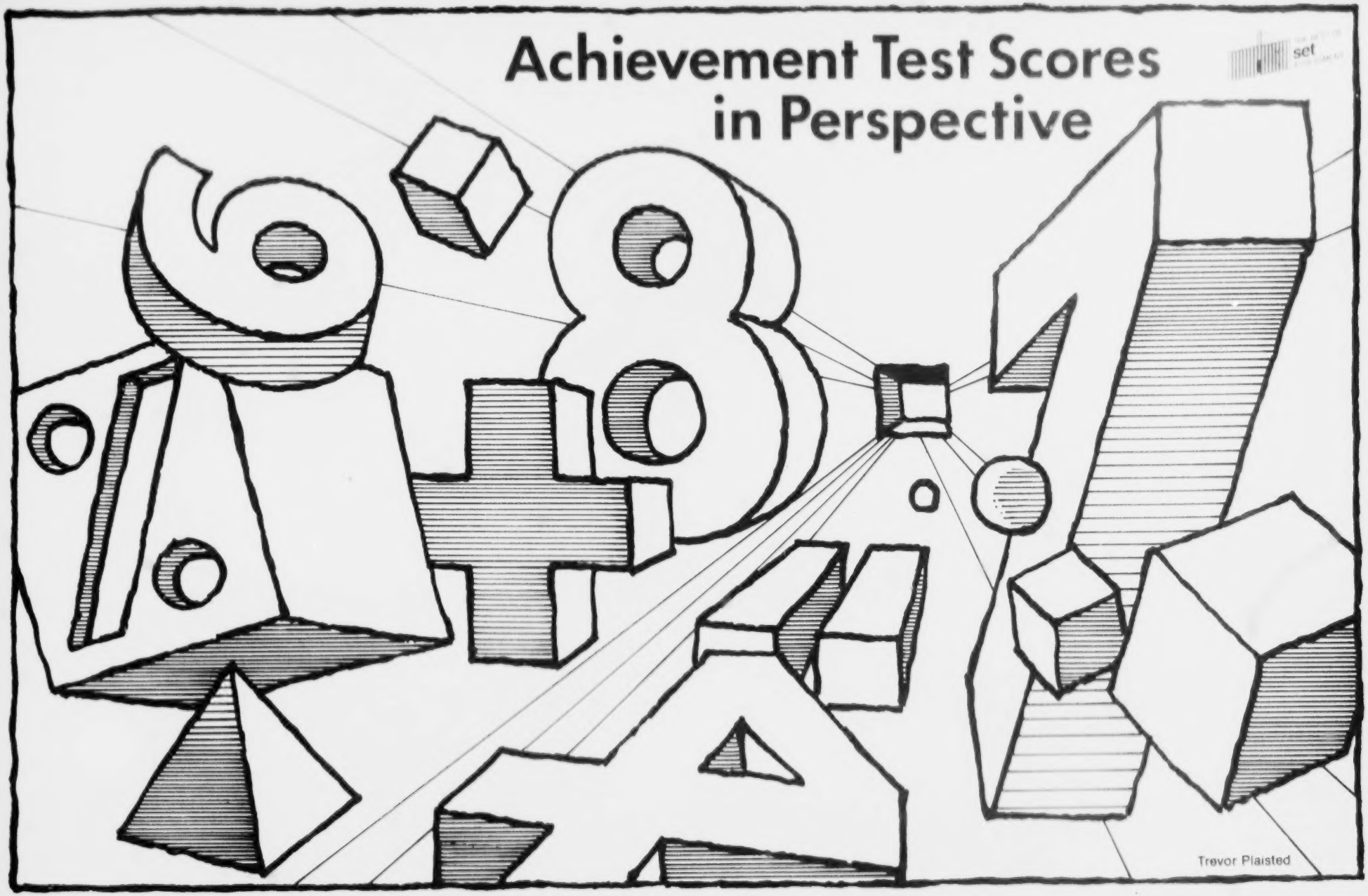
There is also an obligation to provide assessments that are both criterion-referenced and norm-referenced. The criterion-referenced assessments can provide a clear indication of what the student has learned effectively and what is yet to be mastered. The norm-referenced assessments can offer a point of reference in average performance that can give students and parents some idea of how personal performance is related to general levels of expectation.

To end with a warning. Some teachers prefer to assess performance against the student's 'capacity'. This depends crucially upon the teacher's assessment of capacity. Teachers may have no evidence of capacity apart from the very performances that are the subject of the assessment of achievement. Sensitively used, normative assessments can then provide a useful supplement to criterion-referenced assessments.

Notes
Dr Barry McGaw is Director of the Australian Council for Educational Research, Box 210, Hawthorn, Victoria, 3122, Australia.

Copying Permitted
© Copyright on this item is held by NZCER and ACER who grant to all people actively engaged in education the right to copy it in the interests of better teaching.

Achievement Test Scores in Perspective



Achievement Test Scores in Perspective

By William Turnbull
Educational Testing Service, Princeton

The need to assess how much students have learned has been fundamental in education for as long as there have been students and teachers. Long before standardized tests of achievement came on the scene, teachers were making such judgements. They based them on information gleaned from familiar sources: direct observation of students' work, class recitations, conversations with the student or other teachers, daily quizzes, and final examinations. All these bits of information entered into the marks the teacher gave. They still do and they should.

Since we had all these techniques at our disposal, why then were standardized achievement tests welcomed when they came along? There are several reasons.

An Inexact Business

Teachers knew what an inexact business marking really is, and welcomed a new development that held promise of improving their information. Among the virtues of standardized tests, three were particularly appealing. The first virtue was *accuracy*. Studies were made to compare the amount of random error in the traditional kinds of information with the amount of error in the newer standardized tests in the same subjects, and the findings were consistent: the test scores were consistently more accurate. Furthermore, they had a second, related virtue — *objectivity* — that helped overcome some other problems. Some teachers marked hard and some marked easy, some based their marks heavily on deportment, some on how hard students tried rather than how much they achieved. The standardized tests knew nothing of these things, nor of sex, race, honesty, amiability, or love of one's fellow man; only how much the student had learned. Third, scores on standardized tests of achievement had the virture of *comparability*: they showed not only how well particular students performed in comparison with their classmates but also how well they did in comparison with pupils in other classes, other schools, and other districts.

We had, then, the basis for a fine combination of techniques: standardized tests, which could measure sheer accomplishment in several areas very well, and teacher judgement, which could add dimensions inaccessible to standardized testing but pertinent to the interpretation of pupil scores.

We have learned a hard lesson in the ensuing years: how difficult it it to keep achievement scores in that perspective, to see them as a valuable ingredient in the mix of information that indicates how much an individual student or a group of students has learned in a particular area. Scores are not, by themselves, sufficient. That fact needs to be reiterated. But they do remain a key component in any adequate learning information system.

Three Fallacies

Undoubtedly, testing has suffered more from the excessive expectations of its most devoted advocates than from the attacks of its critics. I have mentioned three virtues of tests that have enhanced their usefulness. Let me balance the account by mentioning three fallacies, all born of over-enthusiasm, that have encouraged the misuse of test scores.

The first I shall call the *Micrometer Fallacy*. Some people have invested test scores with a precision — an infallibility — that they never possessed. The ensuing discovery that they are not perfectly accurate has obscured the fact that they are more accurate than most of the alternatives.

The second error I shall call the *Whole Person Fallacy*: the tendency to read into achievement test scores much more than they really tell, which is simply the amount a student has learned in a given subject. Some people are let down upon discovering that achievement tests fail to measure a variety of traits like honesty or leadership or social consciousness, and in their disappointment over the fact that the tests do not describe the whole person, forget that they do a rather good job of measuring the academic accomplishments they purport to measure.

'. . . testing has suffered . . . from the excessive expectations of its most devoted advocates . . .'

Third, there is the *Equal Preparation Fallacy*. Some people expect the test to compensate somehow for the differences in academic development of children whose learning opportunities have differed dramatically. The test score tells you nothing about the difficulties a student has had to overcome to acquire a given level of proficiency, but it tells you a great deal about what that level is — a fact that is central to deciding what the student is ready to tackle next.

Comparable Results

The same three virtues of test scores — accuracy, objectivity, and comparability — that had appealed to teachers interested in the accomplishments of individual students also commended themselves to administrators. For the administrators the greatest of the virtues was (and is) comparability.

They saw at once the power of standardized tests to permit comparisons of the learning achieved by pupils of different teachers, or in different schools, or in different districts. Since it was well known that some schools were much more demanding than others in their curriculum and in their pass marks, here at last was a chance to put all students on the same footing in an objective comparison of results across schools — to make sure the children were learning as much as they should, as determined by how well the children elsewhere were doing.

In the main, these aspirations were sound. And when the three fallacies have been resisted — when people have not expected the precision of the micrometer, have not looked to achievement tests to measure the whole person, and have not assumed that by standardizing the test you have standardized preparation — then, indeed, the scores have given a new dimension, through their comparability across geographic areas and spans of time, to the information available to educators. It is hard to imagine any other basis we might have had for learning, for example, that the verbal and mathematical skills of students applying to go to American universities have declined in the last 15 years. It may not have been a comforting message, but it was one worth getting.

Let's look at a third major use of standardized achievement test scores in America — in the selection of students, especially for university. This use got an important boost during World War II, when teacher power was unavailable to grade the entrance essay tests. The boards of entry substituted objective tests as a war measure because they could be graded so efficiently, fully expecting to return to the essays after the war. But to the surprise of many, not only were the objective tests easier to grade, the scores on them were at least as effective as the essay test scores had been in measuring the attainments of the students who took them. This conclusion was arrived at first through the general observation of the school and university people involved and then confirmed through careful research. The scores were by and large accurate — more so than scores on the essay tests had been.

'The test score tells you nothing about the difficulties a student has had to overcome . . .'

Students who were admitted proved to have the accomplishments the test scores had promised. The scores provided a common currency, unlike the grade-point averages that reflected difference in grading between schools and between parts of the country. Moreover, they were much less subject to special coaching than the essay tests had been; they reflected more fairly the accomplishments of students from different parts of the country, different schools, different curricula.

Unparalleled Growth

As a result, the selective institutions that used the test scores as one basis for admissions decisions were able to seek out talented students from every corner of the nation and every social stratum. They began to see on campus a more heterogeneous group of students both socially and geographically. Moreover, with objective tests of aptitude and achievement in place and efficient, it was possible to test the enormous postwar wave of college applicants under the new system, and an unparalleled expansion of higher education took place in this country.

The system worked because of the same virtues of standardized tests: accuracy, objectivity, and comparability. Again, however, some people were tempted into entertaining great expectations that could not be and have not been fulfilled. The scores were accurate in the main but not micrometers in their precision, even though their errors were smaller than those of other known techniques. They measured achievement and readiness, but not traits of character and temperament — not the whole person. They made no allowance for the inadequacies of preparation or special circumstances of environment that the student had overcome in school. In short, the assumption of equal preparation remained — and remains — a fallacy.

Guidance for Students

We have discussed three legitimate and important uses of standardized tests: their use by teachers to determine how much the individual student has learned, by administrators to determine how much classes and larger groups have learned, and by university admissions people to discover how well prepared a prospective student may be. Each use has made great contributions to education and society when kept in perspective and used with other pertinent information.

Another important use of test scores is by students — to help them as they examine their educational and career goals and estimate their readiness to

undertake a more or less demanding programme of study as their next step. This use has received less emphasis historically then it should have, but the evidence from a set of tests can add materially to the information available for a student's guidance.

'. . . there are many gifted youngsters who make mediocre records in school but surface through their test scores . . .'

People who now discover that test scores can vary from day to day and from test to test are on the right track. Every examination, every judgement about people, is fallible and has a typical error rate. The standard error of measurement associated with standardized test scores is well known because it is readily determined and regularly announced by the publishers. That does not mean that other forms of measurement, such as essays or interviews or letters of reference or teachers' grades, have no error or a smaller error. In fact, although such errors are rarely reported or even determined, research indicates that, in the typical case, they are much greater than the standard error of measurement of a test score.

Those who call for a high degree of teacher involvement in assessing students in their classes are right, as are those who decry the use of test scores alone as a basis for evaluating the effectiveness of an entire educational programme. There is much more to be considered than is reflected in the score, including the conditions under which the results were achieved. But to say what some recent critics have said — that even if one wants to know how pupils are doing in arithmetic one should be forbidden to give a standardized test in arithmetic — is to swing the pendulum right out of the clock.

Institutions that use test scores to select new students should indeed use other information as well. Achievement tests simply cannot measure the whole person. Obviously, a college should consider the record of previous school performance and the judgements of teachers and counsellors. Yet there are many gifted youngsters who make mediocre records in school but surface through their test scores, as Julian Stanley's work at Johns Hopkins University has amply demonstrated.

Forget the Magic

Finally, those who would ascribe to tests given at school age some magic that enables them to divine genetic intelligence or ability to learn should forget it. Achievement tests measure developed ability — developed in relation to a particular subject or discipline. To a large extent, so do scholastic aptitude tests, although the areas of experience through which a student is prepared for them are much more long term and more pervasive in our society. But equal opportunity in education simply has not been realized. To think that, at 18 years of age, people whose experiences have been vastly different can show their inborn potential through a test of verbal and mathematical reasoning is naive, regardless of their cultural advantages or disadvantages.

Standardized tests of achievement have amply demonstrated their utility over the past two or three generations. Because of their accuracy, their objectivity, and their comparability, they deserve recognition as a powerful tool in education. They have suffered in esteem first at the hands of those who claimed for them a set of qualities they could never attain, and latterly from the protests of those who have proposed that since they are imperfect they be done away with. I suggest we put them in a reasonable perspective as we strive to improve both the tests and their use.

Note

This is the 1978 presidential address of W.W. Turnbull to Educational Testing Service (ETS), slightly edited and printed by permission.

ETS is a private, non-profit, organisation devoted to measurement and research. It was founded in 1947 by the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board. It has its main office in Princeton, New Jersey, and is, in fact, the largest organisation devoted to educational testing in the world.

© Copyright for this item remains with ETS.

The Foundations of School Testing

Dick Frizzell

The Foundations of School Testing

For teachers specialising in assessment techniques, and students anxious to sort out the theory and practice of testing in schools.

By Cedric Croft
NZCER

An understanding of validity, reliability and usability are a must for all test users. The *validity* of a test is an indication of how well it measures what the author claims it will measure; its *reliability* describes the consistency or dependability of its scores; and its *usability* is concerned with its administration, format, interpretation and supply.

An alarm clock that keeps accurate time can be described as being *reliable*, and if the alarm goes off at the right hour the clock is functioning *validly*. If the dial of the clock can be read with little chance of misinterpretation, the alarm control operated readily, and it is robust and easily rewound, the clock could be described as being highly *usable*. If it stops, however, and I fail to reset the hands on winding it up, it is still a reliable clock in that it continues to keep consistent time, but the alarm will not function validly since the time shown on the clock does not conform to standard time. The clock could still be usable, but the ease with which it can be used has been affected by the lack of validity.

Tests also can be highly reliable but not valid for a particular purpose. A diagnostic test of long division for example, can give very reliable results, but it would not be the most valid test to select pupils for an enrichment programme in all branches of mathematics. Test validity is heavily influenced by reliability — the alarm will go at the wrong time if the clock runs slow — but high degrees of validity and reliability alone do not necessarily guarantee usability — think of an accurate but faceless alarm clock. The usability of a test will also suffer if reliability or validity are impaired, for example, if the group the test is to be used on differs substantially from the group the test was developed for.

1. Test Validities

When we have given a test, and have the scores, what may we infer from those test scores? What have we measured? What can the scores tell us? What may we not infer? These are questions about the test's validity. Note that validity is inferred, not measured directly. Evidence of validity is usually presented in a test manual, but validity cannot be regarded as a universal and everlasting feature of a test: it is a quality we must judge; and it may be adequate, marginal, or unsatisfactory for this group, at this time, for this specific purpose.

Up to ten types of validity can be identified but the following four are most relevant to classroom testing: content validity, concurrent validity, predictive validity and construct validity.

(i) Content Validity

You have content validity when the test measures a representative sample of the relevant knowledge skills and behaviour. If fractions are emphasised in your teaching, does the test emphasise them too? If urban drift is not covered in your teaching will its inclusion in an achievement test be legitimate? If drawing inferences was your main thrust in science, how many of the test questions could be answered by recall? Evidence of content validity is crucial when we wish to generalise from an individual's performance on a test to the knowledge and skills the test sampled.

Content validity is of the utmost importance for all classroom tests, and also relevant to behaviour checklists, measures of scholastic aptitude, tests of special abilities, and personality inventories.

For a test to have content validity the test items must measure the behaviour they purport to sample. Descriptions of the course or subject matter, the test objectives, and the nature of the sampling, are critical. Although some objective procedures can be used to help assess content validity, the final judgment must remain the opinion of the test's user and the process will always be largely subjective.

It is always possible that a test regarded as having content validity for one school, could be invalid for another school. This would be the case when these schools had differing objectives, or had chosen different content or had different emphases. A test of reading achievement containing a section on skim reading would be valid for a school that taught the techniques of skimming, for example, but invalid in a school that did not have the development of skim reading skills as one of its objectives. For an achievement test, content validity will exist when there is close agreement between the school's objectives and teaching practices, and the test's coverage. The focus of content validity is firmly on the adequacy of the sampling of course-content, and not just on the appearance of the test. Although a test should look as though it will measure what is claimed, this 'face validity' is not sufficient. Establishing content validity must be a prime consideration for everyone constructing an educational test or examination. But how can it be done? As the first step, draw up a table of specifications showing the weighting and emphasis that will be given to the various aspects of subject-matter and cognitive process. For an example, see table 1.

The essential question to ask is: 'To what extent does the content of this test reflect the knowledge and skills I have tried to develop in these pupils?'

(ii) Concurrent Validity

Concurrent validity is an estimate of the relationship that exists between scores on a test, and some other acceptable criterion. For example, the performance of children on New Zealand's *PAT: Reading Comprehension* might be compared with their performance on the Australian *ACER Paragraph*

Table 1 3B BIOLOGY
TERM 3 Ms BELLMAN

Course Content	Cognitive Process				
	Know-ledge	Compre-hension	Applica-tion	Analysis	Total Items
1. Methods of Science: Testing Hypotheses	4	2	2	2	10
2. Animal Classification	2	4	4	–	10
3. The Plants of the Earth	4	4	2	–	10
4. Populations and Mechanics of Evolution	2	3	2	3	10
5. Evolution, Genetics and the Races of Man	–	3	4	3	10
Total Items	12	16	14	8	50

Reading Test. A high correlation (0.85+) would suggest that these tests are measuring substantially the same skills, so consequently, their concurrent validity is high. Note that this does not shed any light on the nature of the skills being tested. Furthermore, all a low correlation tells us, is that the skills and abilities being sampled by each test differ.

If performance on a test of library skills, for example, was correlated with 'ability to use a library', it might be possible to make a statement about the way in which the test performance relates to 'real life'. However, 'real life' is a difficult thing to measure objectively, and a low co-efficient may just mean that the test is not being compared with a *suitable* criterion.

Essentially, concurrent validity provides confirming evidence of a test's validity — 'validity by association' — the test in question must be valid if it relates well to another that is already regarded as valid. There may be a tendency to overvalue the importance of concurrent validity data because it is numerical, but in reality, concurrent validity provides us with the least information about what a test is actually measuring.

(iii) Predictive Validity

Predictive validity is a measure of the relationship between test scores and some appropriate performance at a later date. In New Zealand it would be possible to investigate the relationship between performance on *PAT: Reading Vocabulary* at Form I, and the marks gained in School Certificate English four years later.

Predictive validity is most crucial for selection and training where a test is being used to forecast likely success in a training programme. In the classroom context, tests of reading readiness or learning disability are examples of tests that must have their predictive validity established.

By and large, teachers need not be concerned about the predictive qualities of standardized achievement tests used in classrooms, since their main concern is with here-and-now performance.

(iv) Construct Validity

Construct validity concerns psychological traits or qualities and attempts to describe the underlying psychological processes that are used in a specific test situation. A psychologist's 'constructs' are similar in nature to a physicist's 'models': both are theoretical notions that are developed to help explain and organise aspects of existing knowledge. Terms such as 'reading readiness', 'anxiety', 'scholastic aptitude, 'critical thinking' and 'reading comprehension' are examples of constructs. The basic question in construct validity is not, 'Does this text measure what the author claims it measures?' but, 'What exactly does this test measure?' The identification of all factors influencing the test score is the aim of construct validation.

Despite the crucial importance of this quality, least progress has been made in establishing sound evidence for the construct validity of most psychological and educational tests. There is no satisfactory single technique for assessing construct validity, nor can it be firmly established by any one study. The methods used to obtain evidence of construct validity include (i) logical analysis of the mental processes used to answer test items (ii) studies of group differences (iii) studies of changes in performance over time, particularly when treatments differ (iv) correlations with other tests (v) intercorrelation of items within the test.

A Final Word on Validity

It is worth stressing at this point that all types of validity are inter-dependent: they provide information on how well the test measures a defined field (content), how it compares with other valid measures of a similar type (concurrent), and how well it predicts future performance (predictive); and all this information may be used when considering the test's construct validity. It is also worth reiterating that a test is not valid or invalid per se: it depends on the use to which it is put.

2. Reliabilities

To be valid a test must be reliable. In fact, test reliability has a ceiling effect on test validity: unless a test measures with some consistency it is not possible to be sure what the test is measuring. The reliability of a test is most often expressed as the correlation between one set of scores (on a test for a specified group) and another set of scores (on an equivalent test for the same group). This correlation, usually called the reliability coefficient, ranges from 0 to 1, which corresponds to a scale from complete unreliability, i.e., a random fluctuation of scores, to complete reliability, i.e., perfect consistency of scores.

Although reliability coefficients of 0.96 or higher are reported occasionally, test constructors are satisfied if they can achieve reliability in the vicinity of 0.90. It is tempting to interpret reliability coefficients as the percentage of scores that are in complete agreement, but this is not correct. However, we can use percentages

to illustrate the relationship between reliability levels and fluctuations in test score. Suppose we divide a class we have tested into two halves on the basis of their scores; then we re-test. How many children will remain in the same half following the re-test? If the test has a reliability coefficient of 1.00 all of them, 100%, will still be in the same half; if the reliability coefficient is 0.96 then 95% will stay in the same half and 5% will have moved from one half to the other; and so on.

Table 2 Interpreting Reliability Coefficients in Terms of Percent of Agreement

Correlation Coefficient	Percent of Agreement by Halves
1.00	100
0.96	95
0.90	90
0.85	87
0.81	85
0.76	83
0.64	80
0.49	74
0.25	66
0.00	50

from Robert L. Ebel, *Essentials of Educational Measurement*, N.Y. Prentice Hall, 1972.

The reliability coefficient gives an indication of whether the test is highly consistent, fairly consistent, or very inconsistent only. However, it is used to determine the 'standard error of measurement', of which more later.

The four major types of reliability that are most relevant to school achievement and aptitude testing are calculated by test-retest, parallel forms, split-half and Kuder-Richardson methods.

(i) Test-Retest

Test-retest reliability is estimated after a test has been given to a group on two separate occasions. The set of scores obtained for each individual on the first administration of the test is correlated with the set of scores obtained on the second administration. This gives us a 'test-retest reliability coefficient'.

What pupils do between the two tests can be crucial. If, for example, they learn things related to the test there can be marked changes in the scores. As a result the test-retest reliability coefficient may be artificially deflated. In addition, doing the test again may not seem a very useful activity from the students' point of view, so the second test may be a much poorer measure than the first. For reasons such as this, studies of test-retest reliability must be carefully controlled if a valid estimate of the test's *stability* over time is to be gained.

(ii) Parallel Forms

In some tests, for example, TOSCA in New Zealand and OTIS Higher in Australia, there are parallel forms of the test, which make it possible to measure the same skills on different test material. Usually a single group does both forms of the test, on consecutive days. Parallel forms reliability is really a measure of the *equivalence* of two forms of a test.

(iii) Split-Half

The practical difficulties associated with test-retest and parallel-forms stimulated the development of alternatives. One of these was to split a test into two reasonably equivalent halves, usually on the basis of odd and even items, so that each subject has a score on the odd items, and another on the even items. The correlation between the scores on the odd and even-numbered items is then calcuated. The split-half technique results in a coefficient of internal consistency that is, in essence, a measure of the homogeneity of the skills that are being tested.

(iv) Kuder-Richardson

Kuder and Richardson developed alternative approaches. Their formula, KR20, which has become widely accepted as a basis for estimating test reliability, requires information on the difficulty (proportion of correct responses) of each item in the test, and on the spread of scores. As the calculation of item difficulties can be a time consuming process, an estimate of a test's reliability can be obtained from Kuder-Richardson formula 21, which is based on the number of items in the test, the mean score and the standard deviation of scores. This 'short cut' approach always gives an under-estimate of the reliability coefficient when the items vary in difficulty, as they nearly always do.

Reliability and Errors of Measurement

A very practical way of thinking about reliability is to consider the extent to which an individual's score may vary from time to time. Every test score is made up of two parts: a 'true score', and an 'error score'. Error scores — which can raise or lower an individual's true score — can come from the test itself, from characteristics of the individual, or from features of the test administration.

If changes in test scores are not large between successive testings, the effects of these 'error variables' have been minimal, and the test is reliable. If a child's score changed from something around the 67th percentile to something around the 33rd percentile for example, over a period of a month or so, and there appeared to be no good reason for the change, the reliability of the assessment would be very much in doubt.

The extent to which an individual's score is likely to differ from the 'true' score, can be calculated and expressed as a 'standard error of measurement'. The standard error of measurement gives us an indication of the absolute accuracy of the test scores and generally speaking the smaller the standard error of measurement, the more reliable the test is. The *Ravens Standard Progressive Matrices* is said to have a standard error of measurement of 3. This suggests that for about 68 percent of cases the errors of measurement will be 3 points or less, but for the remaining 32 percent they will be greater than 3.

The standard error of measurement lets us interpret test scores as a band, rather than a single score. In 68 percent of cases an individual's true score will be + or − one standard error of the raw score. If you score 30 on *Ravens*, there are 68 chances in 100 that your true score

lies between 33 and 27. If the band of scores is broadened to encompass two standard errors, i.e. 24–36, there are 96 chances in 100 that the true score falls within this range. Although it may look as though some precision has been lost, the use of bands of scores increases the chances of reliable measurement.

3. Usability

A test must be suitable for the purposes required, and there are practical questions that must be asked.

(i) Is the test readily available? No matter how valid and reliable a test might be in a particular situation, it will be of little use if you cannot get it. Even good tests date after a time, and they do go out of print. Some tests remain research instruments, and despite sound characteristics may never be published for widespread use. Tests that are published overseas can take up to 20 weeks to arrive in Australia and New Zealand, so supply can be an important factor unless long-term planning is carried out diligently.

(ii) Am I able to administer the test?
Tests vary in their complexity, and hence there are a wide range of practical and theoretical training requirements. There are two questions: 1. 'Do I have the background skills necessary for the competent administration of this test?' 2. 'Am I allowed to administer this test?' In New Zealand NZCER administers a test user qualification scheme. In Australia ACER administers a similar scheme. Certain classes of tests are available only to users who possess recognised minimum qualifications. The test catalogues available from NZCER and ACER list the qualifications and the restricted tests.

(iii) Can I interpret the results?
Test results can be reported in a variety of ways: age percentile ranks, class percentile ranks, deciles, stanines, z-scores, T-scores, deviation, IQs, and so on. You need to be familiar with the properties of the transformed scores to interpret the score. It is also necessary to know what behaviour is being sampled, so that how the test items reflect the abilities or aptitudes being tested can be seen. Good test manuals help.

(iv) How much time will it take?
First, time will be spent on giving the test. Then time must be spent on marking and interpretation. And have you got time to do something about what the test may reveal? Is the ultimate usefulness of the test scores worth the time it will take to get them? For example, the *Doren Diagnostic Reading Test of Word Recognition Skills* is made up of 11 subtests each with at least two sections, and takes three hours to administer. With time for scoring and interpretation to be added, there would need to be very real advantages for the teacher and pupil to justify the time involved.

(v) What will the total cost be?
This also must be weighed against the ultimate usefulness of the information gained. Take the probable effective life of the programme into account, as the setting up costs are relatively high, but costs decrease with subsequent use because most of the components are re-usable.

(vi) From what group were the norms derived?
Much of the value in a standardised test comes from being able to compare an individual's score with those of a representative sample of peers. To be of most value, the norms on a test should be derived from the same population as those taking the test. If this is not the case a judgment must be made whether valid comparisons can be made between those taking the test, and the norms sample. Little purpose would be served in administering a test of mechanical comprehension to third-form technical students if the norms for the test had been derived from the performance of university engineering students.

One result is that teachers and other test users are forced to use unadapted overseas norms when interpreting the performance of New Zealanders and Australians. This is a far from ideal situation, which gives rise to test information that is neither valid nor reliable. The widely used *Burt (Rearranged) Word Reading Test*, and its successor the *Burt Word Reading Test (1974 Revision)* can be used to illustrate this point. The *Burt (Rearranged) Word Reading Test* was normed on a sample of Scottish children in 1955, the arrangement of words in the test and the procedures for computing the so-called 'reading age' being adjusted to reflect the word reading skills of Scottish children. This test has been used in its original form in New Zealand, on the assumption that there is no difference between the performance of New Zealand and Scottish children on the test!

A revision of the 1955 version of this test was published in 1974. The Scottish figures show that children now have to read more words correctly to get the same 'reading age'. For example, on the 1955 formula 30 words read correctly gave a 'reading age' of 8.0 years, but in the 1974 revision, 30 words read correctly result in a reading age of 6.7 years. To be credited with a 'reading age' of 8.0 years, 48 words now need to be read correctly. The *Burt Word Reading Test (1974) Revision* is now being widely used and the assumption is that these new Scottish norms validly represent the performance of New Zealand children on this test, which seems unlikely.

The New Zealand test user is not well supplied with tests which present current New Zealand norms. Apart from the *Progressive Achievement Tests*, locally normed tests which have direct application in schools are restricted to the *Otis Tests of Mental Ability*, the *ACER Silent Reading Tests*, the *ACER Arithmetic Tests* and the *Oral Word Reading Test*. The last three named tests were all standardized in 1954, so the normative data is now somewhat limited. Due in 1981 is the New Zealand developed and normed Test of Scholastic Abilities (TOSCA) which should replace the OTIS. Due also in 1981 are two other normed tests, the Proof Reading Test of Spelling and a New Zealand standardization of the Burt Word Reading Test.

On the surface at least, test users in Australia are

better supplied. They have more achievement tests (e.g., *ACER Primary Reading Survey Tests, ACER Mathematics Profile Series*) more general ability tests (e.g., *ACER Lower Grades Ability Scale. ACER Tests of Learning Ability*) and more special purpose tests (e.g., *ACER Checklist for School Beginners, ACER Shorthand Aptitude Test*). However, the large and diverse population, coupled with differing state educational systems, may mean that the number of tests suitable for a specific use is strictly limited. Tests are usually developed for particular populations and education systems so, the Australian user must be vigilant in making sure that the test is suitable for his or her purposes.

Conclusion

By considering a test's *validity*, *reliability* and *usability* it will be possible to decide whether the test will perform the function you have in mind. If an evaluation of the test indicates that it does not meet your requirements, time, effort and money have been well saved. It will also ensure that tests are used as the servant of teachers and pupils, and do not become their masters.

Evaluation Checklist

The characteristics of tests that have been outlined in this article can be used as a basis for evaluating the potential worth of any test. By systematically recording information about these major characteristics, information that can be used in decision-making can be quickly summarized. Item 9 of this *set* is such a list. It is not copyright and you may duplicate it as you wish.

Suggestions for Further Reading

Bauernfeind, R.H.
Building a School Testing Programme. Boston, Houghton Mifflin, 1963. (A good introductory text.)

Ebel, R.L.
Essentials of Educational Measurement. Englewood Cliffs, New Jersey, Prentice-Hall Inc., 1972.

Gronlund, N.E.
Constructing Achievement Tests. Englewood Cliffs, New Jersey, Prentice-Hall Inc., 1977.

Lyman, H.B.
Test Scores and What They Mean. Englewood Cliffs, New Jersey, Prentice Hall Inc., 2nd ed., 1971. (A good introductory text.)

Thorndike , R.L. and Hagen, Elizabeth
Measurement and Evaluation in Psychology and Education. New York, Wiley, 1969.

Test Evaluation Sheet

Identifying Information

Test Title _____ Author(s) _____

Publisher _____ Publication date _____

	Characteristics	Comments
Aids to Inter-pretation	Functions outlined Guide for interpretation Guide for use of results	
Validity	Type(s) reported and values shown Criteria described Face — item arrangement — page layout — quality of illustrations	Content _____ Concurrent _____ Predictive _____ Construct _____
Reliability	Type(s) reported and values shown Method used Sample(s) described	Split-half _____ Parallel forms _____ KR _____ Test-retest _____ Standard Error of Measurement _____
Usability	Administration — Special training necessary — time – power/speed — suitability of instructions — practice exercises — overlapping /discrete Scoring — type of key — ease of conversion — treatment of errors Examinee appropriateness — instructions — items — mode of response Norms — types of derived scores — age range covered — population described — sampling/number of cases	PRs (age) _____ PRs (class) _____ Deciles _____ IQs _____ Grade equivalents _____ Stanines _____ Other _____
Economy	Cost of each component What is re-usable? Time for test marking Time for interpretation	Manual $ _____ Booklet $ _____ Answer sheet $ _____ Marking Key(s) $ _____ Other $ _____

Test Evaluation Sheet

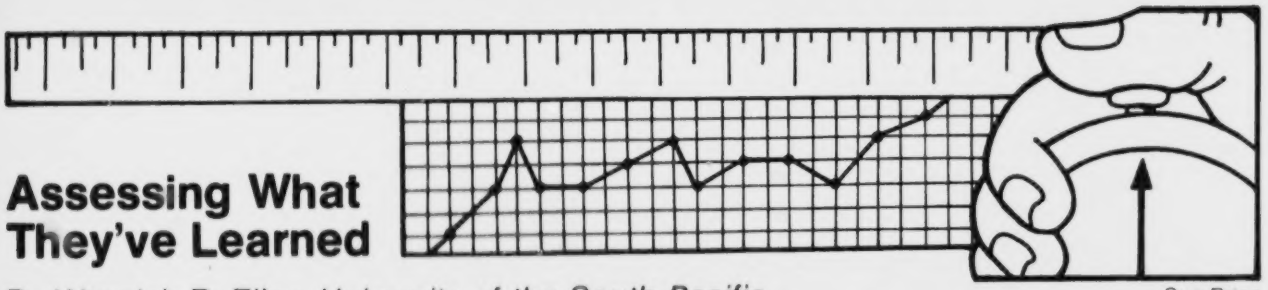
Form devised by Cedric Croft, NZCER

Identifying Information

Test Title: The Larsen-Hammill Test of Written Spelling — Author(s): Stephen Larsen & Donald Hammill

Publisher: Academic Therapy Publications — Publication date: 1978

	Characteristics	Comments
Aids to interpretation	Functions outlined	Two primary functions (assess spelling level and specify areas of weakness) plus four other uses described.
	Guide for interpretation	Interpretation of four patterns of scores given; includes explanation of derived scores.
	Guide for use of results	Some very general guidance in use of results given.
Validity	Type(s) reported and values shown	Content: Achieved by analyzing 10 basal spelling series.
		Concurrent: Range: .69–.92; Med: .82 Predictive: Not reported.
		Construct: Some evidence presented.
	Criteria described	Clear description given of validation criteria.
	Face — item arrangement	N.A.
	— page layout	Answer sheet and class score sheet clearly presented.
	— quality of illustrations	N.A.
Reliability	Type(s) reported and values shown	Split-half: Not reported Parallel forms: 1 form only
		KR21: range: .78–.90; Med: .89 Test-retest: Not reported!
		Standard Error of Measurement: Range 1.4–3.8; Med 3.5
	Method used	Reliability data gathered in conjunction with standardization. Sampling as for norms.
	Sample(s) described	
Usability	Administration	
	— Special training necessary	Not needed for administration.
	— time – power/speed	Time not stated. A power test. Approx. 20 minutes
	— suitability of instructions	General directions only. Precise wording left to examiner
	— practice exercises	None included.
	— overlapping/discrete	Two separate subtests — predictable and unpredictable words. Some overlapping within each.
	Scoring	
	— type of key	List of correct spellings in manual.
	— ease of conversion	Simple conversion – clear tables.
	— treatment of errors	No specific guidance.
	Examinee appropriateness	
	— instructions	Examiner given some discretion.
	— items	Satisfactory – reflect content validity for the U.S.A
	— mode of response	Examinee writes words.
	Norms	
	— types of derived scores	PRs (age): — PRs (class): — Deciles: —
		IQs: — Grade equivalents: ✓
		Stanines: — Other: Splg Ages and Splg Quotients
	— age range covered	Age 5·0–13·5; Grades 1–8 (USA).
	— population described	Described by location, sex, class, age
	— sampling/number of cases	Total sample: 4544.
Economy	Cost of each component	Manual $ 5.00 (US) Booklet $ N.A. Answer sheet $ 1.50 per 25
		Marking Key(s) $ N.A. Other $ 0.20 Class Score Sheet
	What is re-usable?	Manual
	Time for test marking	About 2 mins. per paper.
	Time for interpretation	Estimated as 2 hours per class lot



Assessing What They've Learned

By Warwick B. Elley, *University of the South Pacific*

Sue Price

1. Introduction

Why do we test pupils?

Most teachers spend a lot of time and effort preparing tests, giving tests, marking tests and using tests for one purpose or another. Why? The main functions of tests and examinations in school are:

(a) *Mastery*. Classroom tests are often prepared by teachers to see whether pupils have mastered a particular unit or skill that has just been taught.
e.g., Teacher gives a quick quiz to see whether pupils have learned how to multiply fractions; or know the main events and characters in the book "Animal Farm"; or have learned the main vitamins and the foods that contain them.

(b) *Diagnosis*. Tests are often used by teachers to determine the major weaknesses that a child shows in basic skills.
e.g., In mathematics, to see whether pupils know basic number facts, or can handle zeros, or understand how to divide fractions; in reading, to see whether pupils' difficulty is due to poor vision or hearing, or word attack skills, or vocabulary weakness, or some misconceptions about print.

(c) *Reporting Progress*. Formal examinations are often used to report on the progress made by pupils over a term, or school year. The results are of interest to parents, pupils, potential employers, and teachers in next year's classes.

In addition to these three main uses, formal tests and examinations are often used

(i) to place children in ability groups;
(ii) to select pupils for further education, scholarships, etc.;
(iii) to match pupil materials with pupil abilities;
(iv) to evaluate one's own instruction;
(v) to assist with vocational guidance;
(vi) to determine a child's readiness for learning;
(vii) to undertake research on pupils' abilities, or on teaching methods;
(viii) to determine whether educational standards in a school, or total system are changing.

As teachers, we should be clear *why* we are testing. We should not test just because it is always done. Certainly, tests often do help motivate students to work harder. But the results can be discouraging too, if the results are poor.

In fact, the purpose of the test should affect the kind of test given. Thus, a formal selection examination will normally cover a large number of skills and topics lightly, and most questions should be of middle difficulty level. Mastery tests on a particular unit will be more intensive, with several questions on each of a few topics to see whether they are mastered or not. Some tests should be relatively hard for all pupils, (e.g., Diagnostic tests). Some will be lengthy and formal with elaborate marking schemes (e.g., End-of-year examinations); at other times the teacher will give short, informal quizzes with emphasis on quick feedback. Sometimes the need is for many short questions which can be objectively marked. At other times the teacher's purpose will be better served with a few long answer questions. It is important therefore that we think through our reasons for testing.

2. Qualities of a good examination

Not all examinations are well set. Often they are too hard, or too easy, or not enough time is allowed. Sometimes the questions are vague, or trivial, or provide clues for the 'test-wise' pupil. Many examinations are unbalanced, providing too many questions on some aspects, and too few on others. Such weaknesses means that decisions based on the examination results will be unfair, or pedagogically unsound. How can we tell if our examination is a good one? Two important features are *Reliability* and *Validity*.

(a) *Reliability*

Tests are reliable if they produce *consistent* results, if they produce similar marks on different occasions. If a pupil gains 75%, in a reading comprehension test today, and only 50% tomorrow, then the results are not consistent; the tests are not sufficiently reliable to base judgements on. If a pupil is placed first in his class in a test of multiplication and division of decimals today and is 20th in a similar test tomorrow, we can conclude that the tests are not reliable indicators of his ability.

To be reliable a test must normally be *long enough* to minimize the effects of chance factors in the content and skills included in the test. With a short test, a pupil may be lucky, because he happened to know or guess correctly the few questions that were asked, whereas he knew very little about the areas untouched by the test. A standardized test of reading, mathematics or language normally needs at least 40 good objectively marked questions, to reach a satisfactory level of reliability. To make decisions about individual pupils, for placement, or grouping, or diagnosis, a teacher-made test will probably require more questions than this; for judgements about the performance of whole groups, a teacher can get by with fewer.

Just how long a particular test should be, depends on the type of material tested, the amount of supplementary information available, and the importance of the decisions being made. Thus a test of a highly specific skill, such as arithmetical addition, or spelling, or typing, may produce reliable results within ten minutes. If however, we wish to examine a pupil's grasp of a variety of mathematical relationships, or his understanding of a period of history, and then to make decisions about future schooling on the basis of the results, we may wish to extend the test over two or three hours to gain maximum reliability. For such general skills as essay-writing ability, or oral expression, it is commonly found that pupils vary so much in their performance from day to day and from topic to topic, that the only sure way to gain adequate reliability is to test the

pupils on several occasions and several topics, and combine the marks given by two or three independent markers. This may not always be practicable, but we should realize when our results are likely to be fallible.

Other requirements of a reliable test are clear, precise directions, and reasonable time limits. If students are rushed, their performance may not be typical. The questions should be clear and unambiguous, neither too easy nor too difficult; they should normally discriminate well between good and poor pupils, and they should be capable of reasonably objective marking. Otherwise the results will vary according to the values and whims of the marker. If a choice of questions is allowed reliability usually drops, because markers cannot compare answers so consistently.

These are some of the more important factors in determining how reliable a test will be. It is possible to assess the reliability of a test statistically, but that is a topic for another time.

(b) *Validity*

A good test must be valid. This means that, in addition to measuring a pupil's achievement consistently, it should be *relevant to the main objectives* of the course. It should cover the unit or course adequately, sampling each content area and skill in appropriate proportions. If a teacher knows precisely what his objectives are, he can usually tell, by analysing the questions of a test, whether they conform closely to the objectives he has adopted i.e., whether the test has 'content validity' for his purposes.

To illustrate, a 60-item test of addition in arithmetic may be highly reliable, and yet be quite invalid for measuring achievment in a course of modern mathematics which emphasizes concepts, relationships and reasoning. The objectives of the test do not match the teaching objectives. A 3-hour written examination in manual arts may give reliable results. But if it does not require students to show the actual skills they have learned, it will have poor validity. The students who do well on a written examination may not be those who do well in the practical skills.

Again a test of geography which focusses on factual details about populations, areas, climate, exports, capital cities, and the like would produce irrelevant results for a teacher who stressed broad concepts, generalized skills and underlying relationships. A valid test of such objectives may require novel or fictitious situations on which to base questions so that a pupil can demonstrate that he has attained these objectives, regardless of the particular factual details he has acquired.

Sometimes tests lack validity because of 'cultural bias'. Questions may be unfair because they assume that the pupil has had particular experiences which he has not had, or read books which were not accessible to him, or seen films or T.V. programmes which he has not seen. Sometimes the test may be invalid because the print is illegible, or the diagrams unclear, or the paper inadequately proof-read. Such problems may distract pupils and cause changes in their rank order. Likewise, cheating will result in invalid results. If pupils can copy one another's answers, or gain prior knowledge of the questions the results will not reflect their true grasp of the subject assessed.

To ensure maximum validity for his tests, then, it is important for a teacher to spell out, as clearly as possible, precisely what his objectives are, and to build his questions around these, in the appropriate proportions. Tests which develop without such planning often degenerate into factual quizzes of the low-level, isolated, easily testable fragments of the course.

Guidelines for Checking Reliability

1. Is the test long enough?
2. Are the questions clear?
3. Are the time limits realistic?
4. Are the questions of appropriate difficulty?
5. Is the marking effective?
6. Are the instructions clear?
7. Has the choice of questions been kept to a minimum?

Guidelines for Checking Validity

1. Are the questions relevant, important?
2. Have all topics been assessed in appropriate proportions?
3. Have all skills been assessed in appropriate proportions?
4. Are there clues to the right answers?
5. Is the typing and presentation adequate?
6. Have all students had an adequate opportunity to learn the material tested?
7. Is security adequate to avoid cheating?

3. Planning the test

If a test is not well balanced it will not be valid. Therefore, to ensure proper balance, it is a good idea to draw up a plan or blueprint. List the main topics to be covered on one axis and the major skills to be developed on the other.

Thus, a blueprint for a unit on Mathematics might look like this:

	Skills			
Topics	**Memory**	**Computation**	**Application**	**Total**
Sets	3	—	5	**8**
Fractions	1	3	5	**9**
Measurement	2	3	5	**10**
Decimals	1	4	5	**10**
Statistics	3	5	5	**13**
Total	**10**	**15**	**25**	**50**

The teacher who prepared this plan has clearly decided that the most important objectives in his course are those concerned with applying the skills learnt in new situations, rather than memory work or routine computational skills. Therefore, 25 of the 50 questions are devoted to application. Likewise, statistics is given more weight than the other topics, although all receive some weight.

Planning of this kind should be undertaken in every subject and the weights given should reflect the amount of emphasis given to the topics and skills during the teaching of the unit or course. For example, in Social Studies, the topics to be weighted might be:

Location, Climate, Discovery, Early Settlement, Industry, Transport, Culture. The skills assessed might be Recall, Comprehension, Application, and Evaluation.

An English test might have as its main topics: Written Language, Oral Language, Grammar, Fiction, Poetry. The skills to be tested might be Knowledge, Comprehension, Application, Synthesis (production of original work), and Evaluation.

4. Test questions

Question writing is an art that depends on clear understanding of the subject and of the pupils being assessed, as well as a grasp of the general principles of item writing. It helps, also, to have plenty of time, some imagination, access to other people's questions as models, and an opportunity to have your questions edited by colleagues.

Several kinds of questions can be used, and none is ideal for all circumstances. Written test questions can be simply divided into two types:

(a) *Objective Questions*: These have right or wrong

answers and markers should agree on which are right and which are wrong.

(b) *Subjective Questions*: Essay-type tests in which the pupils must respond to open-ended questions by composing their own answers. There are varying degrees of completeness and correctness.

OBJECTIVE TEST QUESTIONS

(a) *Multiple-choice*

These consist of a *stem*, stating the question, and 4 or 5 possible *options* to choose from.

e.g. (i) What is the area of a rectangle which is 5 cm long and 3 cm wide?

A. 8 sq cm.
*B. 15 sq cm.
C. 16 sq cm.
D. 30 sq. cm.

Note that the "distractors" (A, C and D) should be plausible answers for the pupils who might be unsure.

(ii) If bread is placed in a refrigerator, it will not become mouldy so quickly, because:

*A. cooling slows down the growth of fungi
B. darkness retards the growth of mould
C. cooling prevents the bread from drying out
D. mould requires both heat and light for growth

This question requires the pupil to apply his knowledge of the relationship between temperature and the growth of moulds.

(b) *Matching Questions:*

These consist of two columns of items selected so that pupils can match the words or symbols in one column with the appropriate word or phrase in the other. Matching questions are useful for testing homogeneous sets of facts e.g., matching books with their authors, chemicals with their formulas, words with the parts of speech they represent, etc.

Country	*Capital City*
1. Fiji ( )	A. Rarotonga
2. Tonga ( )	B. Suva
3. Western Samoa ( )	C. Vila
4. Cook Islands ( )	D. Nuku'alofa
	E. Honiara
	F. Apia

Note that both lists should be *homogeneous*, and one should be *longer* than the other. The main fault in preparing these questions is that each list often contains terms which are heterogeneous. They should all be authors, or cities, or chemicals, or parts of speech, etc. If cities are mixed with minerals, and people, and organizations, they provide obvious clues to help the uninformed pupil.

(c) *True-False Questions:*

These consist of a single statement which the pupils are to mark *true or false*, or right or wrong. They are useful questions for a quick quiz, but guessing can be a serious problem with this type of question. This can be reduced somewhat by asking pupils to correct the false statements. e.g.,

		Ring T or F	If F write the correct answer
(i)	25% of 44 is 4	T or F	_____
(ii)	The volume of a mass of gas tends to increase as its temperature increases	T or F	_____
(iii)	Fiji's chief export is copra	T or F	_____

(d) *Completion Questions:*

These consist of a question or sentence containing a blank, for which the pupils must supply the appropriate word, symbol or phrase. These questions are actually "semi-objective", because there is often more than one acceptable answer.

e.g., (i) What is the name of the instrument used to measure temperature? _____

(ii) The device used to tell whether an electric charge is positive or negative is: _____

(iii) What is 25% of 44? _____

Which Type of Objective Test Question Should You Use?

There is *no one best* type of item. All are appropriate at one time or another, but multiple-choice questions are more widely used than others in standardized tests and important examinations. The following advantages are often claimed for multiple-choice questions.

(i) They are more objective and reliable than essay tests or completion questions.
(ii) They make possible the testing of a larger sample of the pupils' knowledge and ability in a short time than does the essay test.
(iii) They enable the teacher to measure process skills as well as recall of simple knowledge. By contrast, true-false, matching and completion questions are largely restricted to simple recall.
(iv) They are easy to mark in large numbers.
(v) They make it impossible for a pupil to gain a high score by guessing.
(vi) Common weaknesses in pupil knowledge and skills can be readily diagnosed by the teacher.
(vii) The questions themselves can be readily evaluated and improved by means of item analysis.

On the other hand, multiple-choice questions do have these disadvantages:

(i) They cannot measure pupils' creative skills, or ability to organize material in a coherent manner. This is particularly important in language, literature, and other expressive subjects.
(ii) They take much time and skill to construct. Poorly prepared questions may produce more invalid results than completion or essay questions.

5. Suggestions for preparing test questions

1. *Essay Questions:*

(i) Specify clearly what is to be included in the answer.

Compare: (Poor) : Write an essay on the French Revolution.

(Better): In not more than 500 words,
(a) Outline the main causes of the French Revolution.
and
(b) Explain why reform could not be obtained without violence.

(ii) Use several short questions rather than one long one.
(iii) Avoid optional questions where possible, as they make marking more difficult.
(iv) Before the test, prepare a model answer, outlining the main criteria and weights to be attached to each.
(v) Mark one question for all pupils before beginning the next.
(vi) Mark without knowing the pupils' names, where possible.
(vii) Obtain independent assessments, wherever you can. The average of two markers is more reliable than the results from one.

2. *Objective Questions:*

(A) GENERAL

(i) Keep your questions brief, simple in expression, and free from complex verbal instructions, double negatives, etc.
(ii) Test only the important facts and skills. Avoid trivial questions, "catch" questions, and irrelevant material.

(B) MULTIPLE-CHOICE QUESTIONS

(i) The problem should be clearly stated in the stem of the question.
e.g., **(poor)**
Bats
A. drive off harmful birds
B. are enemies of man
*C. eat insects
D. eat rats
The pupils must read all options before they understand what the problem is.

(ii) Use only plausible distractors
e.g., **(poor)**
The Prime minister of Fiji is
A. Mr Muldoon
B. Mr Fraser
*C. Sir Kamisese Mara
D. The Shah of Iran
Many pupils could guess the answer with very limited knowledge. Names of other prominent Fijians would provide better distractors.

(iii) Ensure that there is only one correct answer
e.g., **(poor)**
The population of Hamilton is
A. less than 50 000
B. between 50 000 and 70 000
C. over 70 000
D. over 80 000
Both C and D are correct.

(iv) Avoid the stereotyped language of textbooks in the correct answer
e.g., **(poor)**
The Renaissance in Europe was characterised by
A. a decline in trade
B. many religious wars
*C. an unusual efflorescence of creative talent
D. the loss of colonies

(v) Beware of grammatical clues and verbal associations
e.g., **(poor)**
The French scientist who discovered the basis for pasteurising milk was
*A. Louis Pasteur
B. Isaac Newton
C. Francis Bacon
D. Alexander Graham Bell
There are two clues to the right answer here. The question should be rephrased.

(vi) Make the correct option the same length as the distractors
e.g., **(poor)**
Sweets are not recommended for eating between meals as they
A. cause diabetes
B. supply excess energy
C. stimulate the bile
*D. dull the appetite for foods rich in other necessary elements
D sounds right because it is longer, and so makes for a fuller statement.

(C) COMPLETION QUESTIONS

(i) Use a single blank in each question
e.g., **(poor)**
The "_____ of _____" was written by _____.
A pupil may know the facts required, but be confused by the question.

(ii) Place the blanks near the end of the sentence
e.g., **(poor)**
_____ is the name usually given to the breakdown of the soil by various processes.
e.g., **(better)**
The breakdown of the soil by various processes is usually called _____.

(iii) Make all blanks the same length
e.g., **(poor)**
Villa is the capital city of the ___ _____.
A pupil who was not sure whether to choose Solomon Islands or New Hebrides would have an obvious clue here.

(iv) Make sure that there are a finite number of acceptable answers
e.g., **(poor)**
Columbus discovered America in _____.
e.g., **(better)**
In which year did Columbus discover America?
_____.

6. Conclusions

Much more could be said about writing sound questions. However, a careful reading of the principles outlined above and some meticulous editing by your colleagues should make for better reliability and validity than that in a test which grows 'Topsy-like' without planning and forethought.

Teachers who wish to improve their assessment skills further can learn much from studying examples of well prepared examinations and standardised tests, and from analysing the results of their own tests, using item analysis. This and other topics can be followed up in such books as:

Ebel, Robert L., *Essentials of Educational Measurement*, New Jersey, Prentice-Hall, 1972 (the test developer's bible).

Peddie, Bill, and Graham White *Testing in Practice*, Auckland, Heinemann Education, 1978, (short, pithy, a practicing teacher's guide).

Queensland Department of Education, *School Assessment Procedures* Titles: 1 An Introduction, 2 The Multiple Choice Item, 3 Assessment in English, 4 Moderation Within Schools, 5 Assessment in English, 6 Assessment in Foreign Languages, 7 Planning a Summative Assessment Programme, 1971–5. Available from ACER.

Izard, J.F. *Construction and Analysis of Classroom Tests*, Melbourne, ACER, 1977.

Criterion-referenced Measurement

Dick Frizzell

Criterion-referenced Measurement

Glenn Rowley and Colin Macpherson
Monash and La Trobe Universities

A man once owned a dog which was inclined to jump over the back fence and enjoy the delights of the neighbourhood. Deciding that he needed a new fence around his yard, the man was confronted with the problem of determining how high the fence should be. Because he wanted to approach the task systematically, he took his dog to a testing agency, where the animal was put through an extensive series of jumping tests. Eagerly, he awaited the results of the tests, which took some time to arrive. The testing agency you see, carried out a nationwide dog-testing program, and the results had to be processed by computer, along with those of thousands of other dogs.

Finally, the test results arrived in the mail. They were very detailed. His dog, he learned, was about average for Australian dogs. It was, however, well above average for daschunds, and a little below average for greyhounds. He was told that nationwide norms and even neighbourhood norms could be provided given time and money. Although he'd love to have known how his dog compared with these in the next street, the owner regretfully declined. The tests, unfortunately, had not told him what he had set out to find — how high a fence his dog could jump. Had he asked an unanswerable question or had he just asked it of the wrong people?

For twenty years now protagonists of criterion-referenced measurement have been saying that testing has been giving us the wrong kind of information. Tests, they argue, have been designed to provide *relative* information about children (where does Johnny rank), when what we need is *absolute* information (what skills does Johnny possess?). Since about 1970, considerable effort and scholarship have been dedicated to finding, developing and promoting ways to make tests which yield this latter sort of information. These efforts have resulted in new terminology, new ways of constructing and analysing tests, and new ways of reporting, explaining and interpreting children's performances.

What is criterion-referenced measurement?

If I were to tell you that Mark had just scored 15 out of 20 in a geography test, what would you know about him? Very little! Geography tests cover a wide range of content, and even those written for one specific grade level can range from very easy to very difficult. For many tests, his score would be dependent as much on the judgements and whims of the marker as on Mark's own performance. If you are to understand anything at all about his achievement in geography, I will need to provide you with more information than just his score.

Historically, educators have recognized that more information is needed, and have sought to provide that information in the form of comparisons. If we knew that 15 was the third-highest score obtained in a class of 31, we would feel more comfortable about evaluating Mark's achievement. If we knew that the average score of the class was 17, we would still feel comfortable about it, although our evaluation would be quite different. If we know about how the class compares with all other classes we will be even more comfortable.

Some teachers have been inclined to treat test scores as if they have absolute meaning — i.e., a score of 80 percent has its own intrinsic meaning, and if Chris scores 65 percent on a History test and 80 percent on a Spelling test, then he did better at Spelling than he did at History. We know, of course, that this need not be so. We know of the differences that exist between teachers in the standards they expect, in the difficulty of the tests that they set, and in the stringency of their marking procedures.

Norm-referencing is one way in which educators have sought to escape from this dilemma. Over the past 70 years or so, an armoury of techniques has been developed that can add meaning to a single test score by comparing it to some reference group (or *norm* group). Thus, standardized tests provide tables, which can be used to refer a single score to the distribution of scores obtained by carefully-chosen representative samples of pupils of the same age or class level throughout the state, or even the nation. Sally's score of 32 in Spelling takes on a new meaning if we know that 75 percent of her peers scored 32 or less, that is, that she is at the seventy-fifth percentile of a national sample of children of her age. Furthermore, it is possible to refer scores from different spelling tests to the same scale, provided the tests are normed on the same group. Norm-referencing has provided us with a range of techniques and derived scores, such as percentile ranks, age and class norms, profiles, standard scores, T-scores, stanines, etc., which are intended to add meaning to a single score by locating it in a distribution of scores from comparable children. It should be remembered, though, that the meaning which is added is *relative* meaning, and psychometricians have sometimes given the impression that test scores have no meaning except relative meaning. One is reputed, when asked, 'How's your wife?', to have replied 'Compared to what?'.

In 1963, Robert Glaser published an article called "Instructional technology and the measurement of learning outcomes" in *American Psychologist*. Although brief, the article proved to be of historic importance, because it was in this article that the term 'criterion-referenced measurement' was introduced to the world. Glaser was deeply involved in the development of procedures for individualized instruction, and he noted that the measurement techniques which he had learned (essentially how to construct and evaluate good norm-referenced tests) did not seem appropriate to his needs. He knew how to build tests that

were effective at spreading kids' scores out, so that an accurate and reliable ranking of their levels of achievement could be obtained. But he wanted a test which could tell him that Penny had effectively mastered this unit of instruction and was ready to proceed to the next. Where Penny stood in relation to other children (or, equivalently, where other children stood in relation to Penny) was irrelevant to the decision which had to be made. What was needed, Glaser argued, was a criterion-referenced test; one which drew its meaning not from the relation between a score and a set of other scores, but from the relation between the test and a criterion (or domain of behaviours) which the test is designed to represent.

Glaser's argument struck a responsive chord among many educators, although his ideas took some years to take root. The real breakthrough began in 1969 when James Popham and Ted Husek published an article in the *Journal of Educational Measurement* entitled 'Implications of criterion-referenced measurement.' Chatty and easy to read, yet bursting with important ideas, this article brought Glaser's concerns before educators in a way which could not be ignored. By raising a host of questions to which adequate answers simply did not exist, Popham and Husek stimulated an explosion of activity in the field of educational measurement which led to the publication of over 600 articles on criterion-referenced measurement by 1978, and which has continued unabated to this day.

Where are we in testing today?

Over the past sixty years, there have been tremendous technological and theoretical developments in testing, particularly the analysis and selection of items, and the refinement of tests. The influence of the computer has been substantial. It has enabled test publishers to run trials on items, and to develop norms on samples running to hundreds of thousands in some cases, with very little inconvenience to themselves. We have seen the development of whole new areas of theory, e.g., reliability, validity, and generalizability, leading to important new ways to appraise tests, and to an awareness of the concept of error of measurement and the lack of precision in all test scores. We have methods of item analysis which enable us to try items out, analyse the responses, and to select and modify items so as to produce a final version of a test having the 'very best' psychometric qualities. And the state of the art is very advanced indeed. In the 1970 manual for the *California Achievement Tests* (CAT), sixteen reliability coefficients were reported, ranging from .977 to .986 over grades 1 to 12. This is incredibly high. In this regard at least, it seems that test development technology has taken us about as far as it is possible to go.

On the other hand, however, we do not seem to be as well advanced in understanding what it is that we are measuring. In measurement jargon, we have learned to understand *reliability* a great deal better than we understand *validity*. While the reliability of the CAT compares favourably with that of a ruler or a tape measure, the measures themselves do not inspire the same degree of confidence. The difference, of course, is that with the ruler, we understand much better what we are measuring. I know fairly well what a score of 30cm means when I measure with a ruler. But I do not understand so well what a score of so many points means on the CAT, or on other achievement tests. Furthermore, when given an individual's score on a test, it is not necessarily easy to see what should be done for the child. Few test constructors would claim that the score alone will tell you. But it may be what you want.

As professional teachers we will not be using test scores in isolation. Other objective and subjective information will put the test score in perspective. The test score tells us that the child stands high or low in comparison with some other children but, as is well known, children can obtain low scores for a variety of reasons. Not all children who score below the 20th percentile on a reading test are the same and nor should we conclude that they should be treated in the same way. If they are treated identically, it is certain that they will respond very differently.

In summary then, during the course of this century, we have become more and more proficient at developing precise measuring instruments, but we have not progressed to anything like the same extent in understanding the measurements we make, or in making use of the information that they give us. Why is this so?

Testing for Competence

The use of tests as devices to certify that certain people have achieved competence in certain fields has a long history, going back at least 4,000 years. Certainly, formal examinations were used in China as far back as 2200 B.C. Public officials at that time were required to present themselves for an examination every three years to determine their fitness to remain in office. If, after three examinations, they could not be promoted, they had to be dimissed! Civil service examinations lasted in China until 1905, and markedly influenced the development of civil service examination systems in Britain, France and the U.S.A.

Examinations have also been used in universities for many years. There are records of examinations being held at the University of Bologna in 1219 A.D. By the middle of the nineteenth century, written examinations were widely used in Britain, Europe and the U.S.A. both for the awarding of degrees, and for deciding who should be permitted to practise professions such as law, teaching and medicine.

The tradition of using tests as *devices to certify mastery* of some subject-matter and/or skills continues today. We are usually reassured by the knowledge that our physician has passed a long series of examinations, and the same is probably true of lawyers, dentists, pilots, teachers, electricians, plumbers and so on. The possession of a certificate attesting to mastery of an area is seen by society as a way of ensuring a minimal level of competence in various professions and skilled trades.

Testing for Differentiation

There is a second enduring tradition which has contributed to the development of our ideas about testing; this is the tradition associated with the *psychology of individual*

differences. Since the late eighteen hundreds, a major interest in psychology has been in the range of different qualities and abilities which people possess, and in finding ways of identifying and studying these differences. Although it was not the first attempt, the 1905 Binet scale is generally seen as a landmark in the study of individual differences. Binet's test was individually administered. Subsequent years saw major advances in the development of measures which could be administered to groups (i.e., written tests). In the United States, the greatest shot-in-the-arm to the development of group testing techniques was provided by the advent of World War I. Upon the United States' entry into the war, various committees of psychologists were organized to contribute to the war effort. One of these was a committee on the psychological examination of recruits. The 'Army Alpha' test which was developed by this committee was the first group intelligence test to be used on any large scale. It was administered to a million and a quarter men during the war, and was used for selecting men for officer training and so on. It appears to have been regarded as an enormous success.

The use of the Army Alpha test during the war sparked off a long period in the United States when the major thrust of education psychology was towards the development of measures of individual differences — firstly the so-called 'intelligence' tests and then later, school achievement tests of one kind or another. Various technological developments helped to kick things along, including the development of mechanical test-scoring machines, and later, of course, the use of the computer. In the U.S. today, school children are given standardized achievement tests to an extent which would stagger most Australians and New Zealanders. It is a routine part of their schooling, and the basis of a multi-million dollar industry.

It is interesting to note that the technology used in the development of large-scale standardized educational *achievement* tests is essentially the same technology that was used in the development of psychological *intelligence* tests. The tests are intended for wide-scale, perhaps nationwide use, therefore, the items have to be ones which test generalized differences. Items that are specific to this or that curriculum, or to this or that class level are excluded. If you want a test to sell nationally, you make a test of 'reading comprehension', or of 'arithmetical fluency', not one on 'ability to read shop names in High Street', or other specific skills we may find it valuable to teach.

Educational achievement tests which have been developed along these lines can sometimes look very much like psychological aptitude or intelligence tests, and often will have similar properties. What is interesting to note is that the more closely the test approximates to this model, the better it looks psychometrically, that is, to the statisticians who test tests. Thus, as an achievement test is successfully revised, its statistical properties keep improving, and it becomes more and more like an intelligence test, and less and less a measure of the actual content that is taught in school.

Item Analysis for Norm-referenced Tests

There are several reasons why this situation has developed. One is the nature of the procedures of item analysis which are commonly used. These techniques, which are described in detail in most textbooks on educational testing and measurement, are used to identify 'good' and 'bad' items in a test. Their effect, generally, is that items that tend to contribute to a wide 'spread' of scores remain on the test, while those that do not are discarded or modified. By following these procedures, it is possible to produce a test which has the finest psychometric properties.

There are, however, serious problems. Firstly, the effect of the item-selection procedure is to exclude items which are unlike the rest of the items on the test, and to include mostly items which are like the rest. So we end up with a test in which the items all measure pretty much the same thing, as of course the test constructer intends. Psychometricians like to describe the items as being *unidimensional* and the test as having a high degree of *internal consistency*. Tests having this property (which is highly valued in psychometrics) are the most effective in spreading people out along a scale. They are, then, very effective, *norm-referenced* tests. They allow us to rank people on the attribute which they measure with the greatest degree of confidence. But there is a price to be paid for this. What is the attribute measured? Is it what we wanted measured? Is the test valid? Unless great care is taken, we can produce a test which provides a pure measure of a pure attribute, but fails to reflect accurately the various emphases of the curriculum.

Given the procedures used in constructing the test, the result is as near to inevitable as anything can be in education. The procedures were designed originally to develop good norm-referenced tests. They have been borrowed from the procedures used in the psychology of individual differences, where they work very well. Psychologists *want* to measure the underlying trait, whether it be general ability, or any of a variety of special abilities. They *want* a measure which is unidimensional, and hence psychological pure. Generally, they *want* a measure which describes a reasonably stable property of an individual. But the procedures we have borrowed from psychologists have served us less well, since we are usually looking to measure changes we, as teachers, have brought about.

The Use of Tests for Selection

In the early part of the century it was necessary to select those in the primary grades who were most likely to profit from a secondary education. Later on, various external examinations in the senior years of high school have filled a similar function, selecting those likely to profit from tertiary education. In every case, education was seen as a commodity which was to be made available to those who could prove themselves most worthy of it. At a time when there was simply not enough institutionalized education to go around this made sense. Where selection is the ultimate aim, we need a test which can spread people out, and make reliable distinctions among them (i.e., a good norm-referenced test). The tests which we used served this purpose well enough, and were therefore satisfactory in their own terms.

If we accept that one of the present roles of the school is to help each child to learn as well as possible, we need to use tests differently. It is only in university entrance examinations that

selection need be the major issue any longer. At all levels below this, and most above, we ought to be using tests as an instructional device — something to help us do a better job of teaching, not something to help us decide who is worthy of the benefits of our teaching. It is the realization and acceptance of this which has led to the explosion of interest in criterion-referenced testing in recent years.

How are criterion-referenced tests constructed?

From what has been written already, it should be clear that the basic difference between a norm-referenced test and a criterion-referenced test is in the way in which scores are interpreted, rather than in the test itself. It is not possible to pick up a test and identify it as a norm-referenced test, or a criterion-referenced test, just by examining it. Any test can be norm-referenced, although some (e.g., those which produce scores with little variability) may not be very effective norm-referenced tests. But not every test can be criterion-referenced. Unless the test has been constructed with that purpose in mind, it may not be possible to relate test scores to a clearly-defined set of skills or behaviours. Tests which contain collections of items based on fuzzily-defined or undefined objectives cannot yield satisfactory criterion-referenced interpretations. If meaningful interpretation is to be achieved, at least the following requirements must be met.

1. *The objectives of instruction must be clearly defined.* If students' performances are to be described as what they can and cannot do, the objectives must indicate precisely what skills or behaviours are aimed at. In practice this means that all objectives have to be expressed in behavioural terms, i.e., by specifying exactly what the students will be able to perform, at the completion of the teaching.

2. *For each objective, sufficient items must be written* to give some assurance that achievement of that objective is being reliably measured. Strictly speaking, one criterion-referenced test measures the achievement of one objective, although the term is used loosely to describe collections of items which measure collections of objectives.

3. *Item selection must be on the basis of how well the items reflect the behaviours specified in the objectives.* Selecting items on other bases (e.g., on how well they spread the scores) makes it more difficult to provide criterion-referenced interpretations of test performance.

4. *Standards of performance must be specified.* Sometimes standards of performance can be related to out-of-school situations. They help define what appropriate standards are. Examples are: the proficiency needed to operate an automobile, the proficiency needed for a particular job (e.g., typing skills), the proficiency needed to be self-sufficient in a complex society (e.g., writing a letter). For other situations, the best we can do is insist upon *mastery*. But what is mastery? Does it require 100 percent success on items relating to that objective? If not, then what level of success do we set as an indication of mastery? Although much work has been done on 'the standard-setting problem' it remains a matter which can only be resolved by the use of human (and, in a sense, arbitrary) judgment.

If these steps have been followed, scores from the test will yield the kinds of information we seek. More detailed and elaborate blueprints for the construction of criterion-referenced tests can be found in other sources, e.g., W.J. Popham's 1978 testbook, *Criterion-referenced Measurement*.

What are the limitations of criterion-referenced measurement?

Naturally there are many, and we can only focus on a few of them.

Criterion-referenced measurement has found its most frequent use with curricula that be defined in a finite number of specific skills or behaviours which the pupils are to master. While for many curricula, this may be possible, it is clearly not universally so. For many teachers, the specification of precise outcomes (and the same outcomes for all pupils) may seem quite incompatible with their approach to teaching. Some teachers may find that some of what they teach is amenable to this approach, and some is not. In this case, their testing strategies might embrace criterion-referenced measurement only in part, and retain more traditional approaches where they seem appropriate.

Some educators have suggested that criterion-referenced measurement is appropriate for assessing the effects of training, as distinct from education. It is possible to distinguish between two types of objectives: *mastery*, or minimum essentials (certain specific skills which can and should be achieved by virtually all students and which are necessary for further study), and *developmental* (more generalized abilities such as problem-solving and clear thinking, which one can never really claim to have achieved, but towards which we hope all our students are progressing). For mastery objectives, a criterion-referenced approach is possible, and probably essential; for developmental objectives it is much more difficult to apply. Thus criterion-referenced measurement has been applied most effectively in the basic skills areas, less so in parts of the curriculum where the air is more rarified and the objectives harder to define.

The focus of criterion-referenced measurement is on the achievement or non-achievement of certain competencies, and the emphasis is not usually on the extent to which a student has achieved excellence in an area. In fact, advocates of criterion-referenced measurement frequently see the task of testing as being to distinguish between students who have mastered an objective and those who have not — between 'masters' and 'non-masters'. But not all of our teaching is of this nature, and teaching which is designed to encourage excellence in a field of study may not fit very comfortably within such a framework. For many, and probably most of the skills taught in schools, there are not just two levels of competence, but an infinite variety, ranging from the highest level of skill all the way down to complete ineptitude. Criterion-referenced measurement, when used to classify pupils into the categories of 'master' and 'non-master',

cannot portray the range of abilities present in a normal group of children.

How long does a criterion-referenced test have to be to provide an adequate sample of a behavioural domain? Rules of thumb do exist (e.g., Popham suggests a minimum of 10 items) but the question is one which admits of no single answer. If the tasks in the domain (and hence in the items) are very similar, we can make do with fewer items; if they are varied, we would need more items to achieve the same accuracy. And, most importantly, the length of the test must reflect the importance of the decisions made from it, and the consequences of being wrong. The more crucial the decision, the more items we would want to include.

Conclusion

The advent (really, the rediscovery) of criterion-referenced measurement has undoubtedly been an important step in our thinking about education and testing. In many ways, we will never be the same again — particularly in the way we report information to parents. To the extent that criterion-referenced measurement has forced us to focus our attention on reporting what children can do and what they cannot do, its effects cannot be anything but beneficial.

But criterion-referenced measurement is not going to be the answer to all our problems, and its advocates would do well to recognize that there are situations in teaching for which it is just not useful or practical. One approach emphasizes the kind of information we get by examining the content of the test itself, the skills required to do well on it, and so on. The other asks 'how well do comparable children do on the same test?' The two kinds of information complement one another, and in most situations we do not have to choose between them — both are useful if they help us to evaluate the child's performance. Both can be obtained by studying the scores from a well-constructed criterion-referenced test. Both can be obtained from a well-constructed norm-referenced test that details item content and has a comprehensive teachers' manual. There may be occasions when we need only information about one child's progress on one skill — a criterion-referenced test is the answer. However, there will be occasions when we want to rank students not on a specific skill but on a broad range of capabilities. In such cases, the traditional type of norm-referenced test would be more appropriate. The type of information that is required and the ways that test scores are going to be used should determine the type of test that is administered.

It should also be pointed out that there is a price to be paid by teachers who want to reap the rich educational harvest offered by criterion-referenced tests. Such testing, if done properly, takes quite a deal of time and effort. Teachers who are already overburdened with the many demands of preparation, teaching, counselling, etc., may find the prospect of preparing whole sets of criterion-referenced tests more than a little daunting, in spite of the benefits to be gained. However, if teachers with similar teaching objectives are willing to share the tests they write, a wealth of criterion-referenced measurement material could be made widely available after only a moderate effort by each individual.

An exciting possibility for the very near future marries the growing interest in criterion-referenced measurement with the introduction of microcomputers into many schools. It appears that the next step to be taken in the computer education movements in many countries may be the linking of individual school systems to regional or statewide networks. (Indeed, this has been the case in Tasmania for some years now.) Criterion-referenced test specifications, by definition, have a very high level of descriptive clarity. Anyone reading them can fairly quickly be made aware of *exactly* what is being measured by a particular test. Imagine the quite feasible situation where a teacher is looking for a test relating to the policies of early colonial governors. (S)he sits down at a school micro-computer, hooks into the network and within minutes is perusing the first level of specifications for criterion-referenced tests that are in some way connected to the keywords (s)he typed into the system. Those tests that look most promising can be evaluated further by calling up the next level of detail in the specifications. Finally, a test that will suit the teacher's purposes is found and at the push of a button the actual test and its specifications are printed out on the school's printer. All that need be expected of the teacher is that (s)he will at some time contribute to this criterion-referenced test bank. But first it would benefit many people if teachers, once they construct a criterion-referenced test, let others in the same teaching area be aware of its existence and availability.

Notes

Excellent elementary level accounts of criterion-referenced measurement are contained in

Popham, W.J. *Modern Educational Measurement*, Englewood Cliffs NJ, Prentice-Hall, 1981 (Chapter 2).

and

Popham, W.J. *Criterion-referenced Measurement*, Englewood Cliffs NJ, Prentice-Hall, 1978.

Two articles of historic importance, and which make excellent reading, are

Glaser, R. 'Instructional Technology and the Measurement of Learning Outcomes', *American Psychologist*, Vol. 18, pp. 519-521, 1963.

and

Popham, W.J. and Husek, T. 'Implications of Criterion-referenced Measurement, *Journal of Educational Measurement*, Vol. 6, pp. 1-10, 1969.

For an up-to-date review of the many recent technical developments in criterion-referenced measurement you might consult either

Berk, R.E. (ed.) *Criterion-referenced Measurement: The State of the Art*, Baltimore, John Hopkins University Press, 1981.

or

Hambleton, R.K. (ed.) 'Contributions to Criterion-referenced Testing Terminology', Special issue of *Applied Psychological Measurement*, Vol. 4, No. 4, 1980.

Dr Glenn Rowley is Senior Lecturer in Education at Monash University, Melbourne; Colin Macpherson is a teacher, and a graduate student at La Trobe University, Melbourne.

set research information for teachers
number two 1987

Investing in Item Banks

By Neil Reid
NZCER

MID-YEAR exams are looming, and Mr Davey has a paper to set for his senior maths class. Going to his classroom store cupboard he drags down a battered manilla folder labelled 'Exams', bulging with dog-eared and yellowing papers. He flicks through the top few copies looking for last year's senior mid-year exam and the ones for the three years previous. On a sheet of lined paper he copies out those questions with double ticks or SSFG (sorted sheep from goats) written in the margin. He studiously avoids those with large crosses alongside or marginal notes of 'hopeless', 'too hard', 'diagram problem', and 'takes too long'. In 30 minutes he had his mid-year exam ready to take along to the school secretary for typing. Mr Davey has, in fact, been using his own embryonic, and rather crude item bank.

What is an item bank?

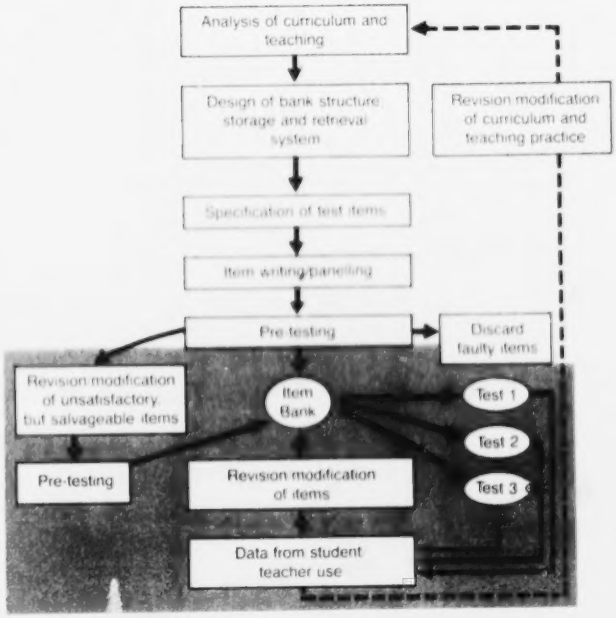
ITEM BANKS, sometimes called 'item pools' 'question banks', 'item files', 'test item libraries' or 'item collections', are variously defined. For the purposes of this article they are regarded as being a *large collection of accessible test questions*. By 'large', we mean that the number of items is many times more than would be used in a single test. 'Accessible' means that the items are classified, indexed, organized or arranged in such a way that they can be retrieved readily for test or exam assembly purposes; there is a system to make it easy for potential users to reference the items and to choose among them. And, under this relatively unrestricted definition, a variety of items (questions) can be considered for inclusion in a bank: true-false, multiple-choice, short answer, extended answer (essay), even practical exercises. The items may cover many topics, different achievement and ability dimensions, and be used for a variety of purposes and with different student groups. They may be classified tightly or be relatively independent of any subject or skill taxonomy.

Any items incorporated in a bank should have been tried out on students and found acceptable; they will be of proven quality. They should also have descriptive and statistical information detailing certain important properties and characteristics (see Appendix).

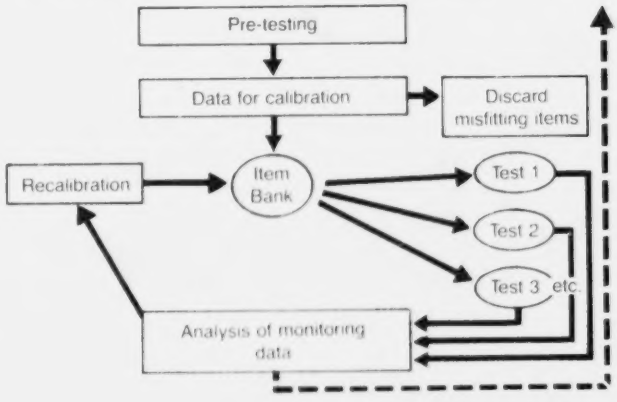
How are item banks developed?

THE SYSTEMATIC DEVELOPMENT of item banks has been attempted in many countries since the late 60s when the pioneering work of Wood and Skurnik in developing a mathematics item bank was undertaken at the National Foundation for Educational Research (NFER) in England. Different strategies have been adopted by different developers, but experience has shown that the steps in the flow chart provide the essential sequence of an item bank's development.

Flow Chart: How to Build an Item Bank



For a calibrated bank the following stages would be substituted in the screened area of the diagram.



Clearly, a bank of this kind is **not** the work of one developer. It does **not** represent one person's notion of what the bank should contain in terms of content, the levels of cognition to be tested, or the mode of testing. Instead, the item collection should represent the consensus of knowledgeable professionals: practising teachers, curriculum specialists, advisers and, for some types of assessment, educational psychologists.

What kinds of items should be in a bank?

AS THE definition suggests, **any** kind of test question can be incorporated in an item bank. Multiple-choice items have tended to outnumber other types in established banks primarily because they are objective, require precise responses, are simple to modify, are easier to validate, and fit readily into most classification systems. They are admirably suited to testing various aspects of mathematics and science, the two subject areas for which most item banks have been developed. However, banks have been developed for the social sciences, too, and for subject areas with practical components such as woodwork and homecraft.

Open-ended questions, like those requiring paragraph answers or essays can also be banked, but marking outlines or guides need to accompany them, otherwise idiosyncratic interpretations of what constituted an 'adequate' or 'acceptable' answer would complicate matters. Indeed, Wood has suggested that there is no reason why tasks, '. . . such as oral questions, dictation, musical passages, project topics, practical experiments, and so on, should not be stored, providing some quantitative evaluation of them can be made.'

How many items should an item bank have?

THAT DEPENDS on several things: the bank's intended purpose(s), who is using the bank, whether the bank is computerised or not, and similar considerations. Probably the best rule is, the more items the better, assuming, of course, that all items are of proven quality, are valid in terms of content, and that the classification and retrieval systems are not overwhelmed by sheer numbers of items so that they fail to operate efficiently.

Crude guidelines for the number of items required to make up a bank, as reported by Prosser, are: 10 items for every one that could be used for any one test, and, 50 items for every hour of classroom instruction on a particular topic. Where banks are referenced to stated learning (or instructional) objectives a minimum of five items per objective is suggested.

Having a large item bank means that a user is more likely to find a suitable match between available items and what has been taught, the kind of test, and the level(s) of difficulty required. A well-stocked bank also ensures that items do not become overused, and this to some degree gets around the problem of item security. Security is important when, for example, item banks are used for moderation purposes.

Where items are to be used in diagnosing learning difficulties then many items for each sub-topic or instructional objective are needed. But if a bank is to be tapped for the purpose of making a comprehensive evaluation of a programme, or a system-, state- or nation-wide survey, then fewer items on each topic or objective, but a large number overall, to assess the many different learning outcomes, would be required; in fact a more general bank is needed altogether.

Where do the items to stock a bank come from?

REGRETTABLY, there are really very few sources of sound, high-quality, and content-relevant test items. Twenty or so commercially published sources are currently available, and several of these are of Australasian origin (see list provided). Other sources of items or ideas for writing items are the workbooks or manuals that accompany published textbooks or instructional materials, materials produced by state departments of education, and one-off tests found in university theses and diploma projects.

Obviously, using existing item collections is convenient and cheap, when compared with the cost and effort of starting from scratch in developing one's own item bank. However, the disadvantage of this approach is exactly the same one that turns some teachers off using nationally standardized tests: the content frequently does not meet local curriculum objectives very well, nor do the questions fit local contexts.

Where users decide to write their own items for a bank, and thus increase the likelihood that the resulting item collection will meet local needs and match curriculum content and emphases more precisely, it is instructive to examine already published items from reputable sources (including standardized tests) as exemplars and as profitable starting points for ideas and format. But, let us not mislead by minimizing the magnitude of the task. Many test specialists have commented on the difficulties encountered in developing local item banks.

Paramount is the problem of obtaining enough high quality items that are unambiguous measures of curriculum objectives (other than knowledge or the recall of factual information which are assessed relatively easily). Hard-pressed teachers working alone, or in small groups, rarely have the time, even if they have the skill, to devise the hundreds of questions required that meet the recognized criteria of 'a good item'. And, even where questions **are** produced locally, there is no guarantee that the items written by one teacher will be acceptable to another.

Let it also be remembered that, no matter how sophisticated the classification and retrieval systems, and despite the enlightened application of the latest in electronic gadgetry, **no** bank is better than the individual items that make it up. It is at this fundamental level that creative energy and expertise must be harnessed.

Nevertheless, there is nothing to stop local groups of enthusiastic teachers working co-operatively over a period of time to produce their own item banks, using their own and others' questions. Teachers' involvement and sense of ownership and acceptance of the resulting item bank(s) is important, and the experience of participating in item-writing workshops and in critiquing the efforts of others (commonly called 'item panelling'), will undoubtedly

broaden their perception on how specific skills and knowledge domains may be reliably and validly assessed. Information on item-writing and trial testing, and advice on setting up and maintaining local item banks, is readily available, both in Australia and New Zealand, to those teachers who wish to try their hand.

How are items in a bank classified?

AS MILLMAN and Arter state: 'Classification is the key that unlocks the item bank.' Unless items in a bank can be retrieved quickly and precisely, the system will grind to a halt; frustrated users will shun it. To ensure an efficient, workable system, stored items must be accompanied by adequate descriptive information.

Two classification procedures have been found to work well. One, a fixed category approach, classifies items by content, often with sub-divisions into topics, subject matter or objectives.

```
30000  ELECTRICITY, ELECTRONICS, AND MAGNETISM
       33000  Magnetism
              33100  theories of magnetism
              33200  magnetic materials
                     33210  ferro-magnetism
              33300  types of magnets
                     33310  permanent
                            33311  bar magnet
                            33312  horseshoe magnet
                     33320  temporary
                            33321  induced
                            33322  electro-  (see also 31620)
              33400  properties of magnets
                     33410  magnetic fields
                            33411  lines of
```

A second classification uses a keyword approach to identify items. This system is flexible; it can handle items which span categories; items can be described in great detail and several kinds of information can be classified together – subject matter, topic, ability/process, class levels(s), item setting, etc.

```
WOODWORK; LEVEL III; JOINTS; TENON;STRENGTH OF; KNOW.;SPEC.FACT; EASY
WOODWORK; LEVEL  II; JOINTS; BUTT;EDGE/CHARACT. OF; COMP.;INTERP.; DIFFICULT
WOODWORK; LEVEL III; JOINTS; DOVETAIL;LAPPED/CONSTRUCT. OF; KNOW.;ROUTINE OP.; MEDIUM
WOODWORK; LEVEL  II; JOINTS; DOWEL;SPEC.USE/SITUATION; ANALYSIS; DIFFICULT
WOODWORK; LEVEL III; JOINTS; MITRE;PICT.FRAMING; COMP.;EXTRAP.; EASY
```

Modification of the system to keep pace with change is also accomplished simply. But, clearly, it requires a relatively powerful computer to operate such a system efficiently and effectively, and this may rule it out for many potential item bank users.

Further details on item descriptors (information and characteristics) for use in banking classification systems, may be found in the Appendix.

Should item banks be computerized?

AUTOMATION makes it feasible to accomplish several important operations relatively easily. Item statistics can be calculated and recorded regularly or at will. Weak items, or those that are seldom used by teachers can be automatically purged. Tests can be assembled by computer to a teacher's exact specifications with considerable savings in time, and, the test text will be free of errors. Adaptive, or tailored testing, even for individual students, is readily handled by computer, as is scoring and fast feedback to both students and teachers via printout. Where a bank is calibrated, other advantages accrue (see below).

While it may appear that a computerized item bank is the answer to every busy teacher's prayer for hassle-free testing, there are disadvantages. Paramount is cost. Hiscox warns us: 'The vision of a general purpose computerized item bank is frequently simplistic or unjustifiable based on the benefits it will produce compared to its cost. There is simply no evidence that a reasonably priced computer system can do all the item banking tasks we would ask of it.' Another is the stark fact that computers, even large and sophisticated machines, cannot handle much of the stimulus material that is so much part of sound testing practice in many subject areas: mathematical and scientific diagrams (more especially those involving diagonals and circles), pictures, line illustrations and cartoons, graphs, special symbols or characters, maps, facsimile texts, and so on.

Already, some banks are available on floppy disk and it appears that very soon items may be able to be accessed by TV/phone. Educational Testing Service in Princeton, USA, reports that its computer systems staff have auto-

mated test development procedures. Such technology will, of course, become available to all in the future. As advances are made in computer technology and good item banking software programs become available for reasonably priced microcomputers, automating item banks at a local level will become an increasingly attractive proposition.

Should item banks be calibrated?

IT DEPENDS on who you ask; it is a 'hot' topic amongst test developers and psychometricians. Proponents of item calibration, using the popular Rasch model or some other model derived from item response theory, will tell you that an uncalibrated bank is next-to-worthless and those responsible for developing it are behind the times and strictly amateurs. But critics who are sceptical of IRT models and the apparent reduction of the rich array of human abilities to homogenous latent traits counter the calibration enthusiast's claims. They say that there is insufficient evidence that item response theory 'works' and believe that the latent trait models generated are simplistic. Special concern is voiced when such theories are applied to achievement testing of the kind that concerns teachers at every level of the education system.

To some extent, whether or not a bank should contain calibrated items depends on the user's purpose(s). Item banks serving classroom teachers' needs exclusively probably do not need to be calibrated. Items for class tests would more than likely be selected to meet subject/topic, ability/process, objectives and/or traditional item difficulty specifications. Statistical criteria, such as those provided for a calibrated item collection, would be of little or no concern. But, if the bank is for subject moderation, or for efficient adaptive testing, or if it is important to have items on a common scale (so that comparisons between students taking different item combinations can be managed) then calibration is obviously essential.

An advantage of calibrating that deserves mention is the technique of sample-free item analysis. New, untried items can be added to a bank without the large-scale pre-testing that is required for conventional, uncalibrated item banking which the flowchart showed.

When new items are to be trialled they are included with already banked items of known characteristics in a test which is then administered. On the basis of the results, the new items' consistency with the bank is evaluated. Where judged satisfactory, the new items can be calibrated onto the bank for later use; any misfitting items weeded out.

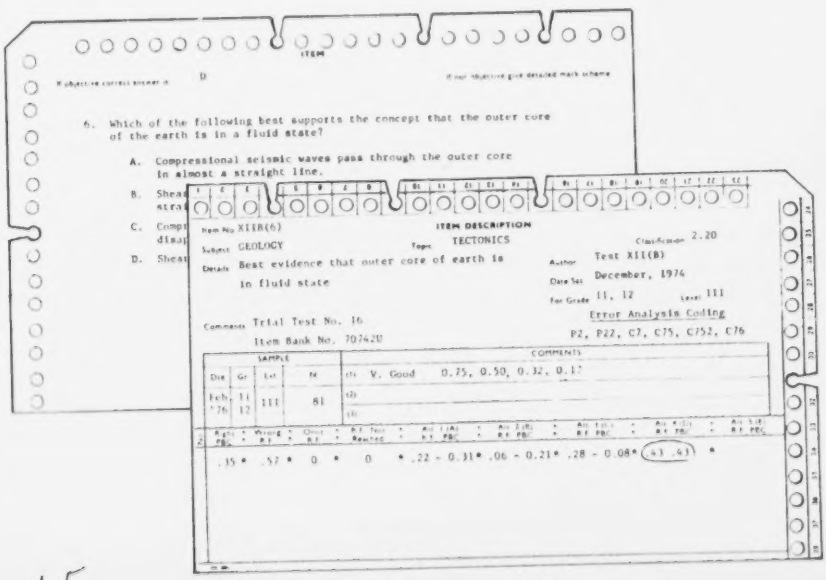
How might an item bank work?

AT a fundamental level, an item bank can be simply a systematized collection of items put together and printed in a booklet or loose-leaf folder, a sort of mail-order catalogue, which is made available to teachers. The teacher who wants to use particular items for a test just copies out what he or she considers appropriate, has it typed up and then reproduced in some convenient way. Better still, whole pages of suitable items direct from the collection are photocopied to make up tests, thus avoiding the introduction of typographical and similar errors and saving the teacher's and typist's valuable time.

An objection to these methods is that the 'capital' of the bank passes out of the hands of the organisers. There is no feedback to them on how items are performing, no provision for up-dating item statistics or for the modification of items. Item security might also become a problem. In fact, such a system is open to abuse once the item bank has been disseminated. And, there is no lack of anecdotal evidence and documented examples of the misuse of published item banks!

A more sophisticated, but also more 'remote' approach involves the teacher filling out a standard form specifying fairly precisely the kinds of items required for a test. The detailed form, a blueprint for the desired test, is sent to the bank organisers in some centralised location who retrieve appropriate items from the bank and compile the test. The teacher using the service is provided with a master copy of the test for reproduction.

At a school or local level, a card index system with items categorized along several dimensions can be operated along similar lines by a teacher or secretary with responsibility for compiling test 'orders'.


ITEM
6. Which of the following best supports the concept that the outer core of the earth is in a fluid state?
A. Compressional seismic waves pass through the outer core in almost a straight line.
ITEM DESCRIPTION
Item No XIIB(6)
Subject GEOLOGY
Topic TECTONICS
Details Best evidence that outer core of earth is in fluid state
Author Test XII(B)
Date Set December, 1974
For Grade 11, 12
Level III
Error Analysis Coding
P2, P22, G7, C75, C752, C76
Comments Trial Test No. 16
Item Bank No. 7074210
SAMPLE
COMMENTS
Feb. '76
V. Good 0.75, 0.50, 0.32, 0.17
81

A combination of these two procedures might have the teacher referring to a master file or catalogue and specifying (by code numbers or keywords) the items arranged in a preferred order to make up a test. The coded information is fed into a computer which prints out high-quality text as a master for cheap reproduction by the teacher.

If the system for compiling tests at a centralized office from 'controlled' item banks also has a scoring service, then most of the objections to 'uncontrolled' published item catalogues can be met. In such a system it will be known which items are being used and by whom, and feedback will be available to monitor item performance and to up-date item statistics. Such a scheme, however, does require knowledgeable staff to operate and maintain the bank and ancillary services, and, it is clearly more expensive to run. But, in these days of 'user pays', it may well be considered a viable operation.

Who would use an item bank and for what?

IT is generally considered that item banks are potentially useful for classroom teachers who: (i) wish to assess their students' learning using measures with known characteristics, (ii) are willing to examine closely what they are teaching and to align their testing with it, (iii) want to save time without sacrificing the quality of their assessments, (iv) are able to appreciate the flexibility of item banks to meet a variety of testing needs – from individualized tests on single sub-topics (to diagnose specific learning problems) to end-of-year surveys of achievement (in one subject area for hundreds of students) and (v) wish to retain control over what is to be tested, how it will be tested, and when.

Others, such as those responsible for conducting or monitoring national examinations, and educational administrators, may also wish to exploit the flexibility and potential of item banking. With 'internal assessment', 'reference tests', 'moderation', and 'school-based achievement', being assessment terms bandied about today, it is not hard to imagine that an informed use of item banking might well assist those concerned with competence, comparability, standards and similar weighty matters. The thought is not new. The Schools Council in Britain (now disbanded) was contemplating using item banking to moderate Mode 3 examinations and to improve the GCE and CSE examinations away back in the sixties. In 1972, Elley and Livingstone, two NZCER research officers, discussed the possibility of item banking as a method of moderation in their publication 'External Examinations and Internal Assessments'. In Australia, ACER began work on item banks in the early seventies, and Tasmania, since 1972 through the Hobart Curriculum Centre under the leadership of Don Palmer, has had centralized item banks for several grade levels in a variety of subjects with clever built-in methods of self-moderation and error-analysis.

What is the future of item banking?

WRITING in 1974, Wood stated: 'Like fume-free cars and the Kingdom of Heaven, question banking is one of those ideas which has great appeal but which people do little about.' He went on to lament the lack of progress following the promising start that had been made in England in developing mathematics item banks.

Thirteen years on: what has changed? Briefly, the many benefits of item banks – principally their flexibility that permits easy adjustment to a variety of instructional/assessment settings – is slowly being recognised by the teaching profession; experimentation with latent trait models has led to more considered and balanced views of the contribution they can make to the item banking enterprise; published item banks have become increasingly available, and, despite some blatant abuses, have found a niche in many teachers' assessment armouries; fears held earlier by some teachers that more assessment and a narrowing of the curriculum to 'measurable outcomes', through the relative ease of testing with item banks, have largely been dispelled – it just hasn't happened; and the impact of computers, of course, cannot be ignored – one can confidently predict exciting developments on this front.

In summary, it would be fair to say that item banking has far from universal acceptance in our schools and other educational contexts. It has considerable unrealized potential, and, optimistically, it **does** have a promising future!

Notes

Mr Neil Reid is Chief Research Officer: Measurement and Evaluation, NZCER, Box 3237, Wellington, New Zealand.

Item Calibration

Item calibration involves evaluating the fit of items to an item response theory model. In the Rasch model it consists of estimating the difficulty parameter value for each item. The great advantage claimed for this particular procedure is that estimates of item difficulty are **independent** of the particular students and other items included in the calibration exercise.

Appendix: Item Information

(Adapted from 'Issues in Item Banking', **Journal of Educational Measurement**, 21:4, 315-330, 1984).

Accurate information about banked items is essential to ensure the efficient operation of any item bank. Depending on the scale and scope of the bank, the following information about **each** item should be considered for entry and retrieval purposes.

Item Description

1. Identification number, sign or symbol.
2. Content/text of item.
3. Keyed answer for objective items; model answer for paragraph/essay questions; typical incorrect responses for diagnostic test items.
4. Required associated stimulus material (graphs, illustrations, diagrams, etc.)
5. Cross-reference to other items or to common stimulus material (reading passage, map, diagram, etc.)
6. Ability/mental process classification.
7. Keyword(s) of item.
8. Author(s) of item.
9. Source of item (published/commercial, Departmental, school, etc.)
10. Revision or version of previous item.
11. Question type (multiple-choice, true-false, essay, etc.)
12. Type of student directions required for item use.
13. Curricular importance (essential, highly desirable, desirable, etc.)
14. Appropriate class/educational level.
15. Cross reference to syllabus, textbooks, teacher's guide, manuals, workbooks, etc.)
16. Security classification (secure, specified use, unrestricted use).
17. Date of item origination.
18. Pre-testing history (date(s), class level(s), number of students, etc.)
19. User comments; suggested modifications.

Item characteristics

1. Difficulty index*.
2. Discrimination index.
3. Item response model fit index (for Rasch-scaled or other IRT calibrated items).
4. Bias index.
5. Readability level index*.
6. Average time for completion.
7. Option response frequencies (particularly for diagnostic tests).
8. Information response frequencies (particularly for diagnostic tests).

* Sometimes judged rather than calculated. In such instances, words (e.g., high, low, hard, easy), rather than figures should be used for these estimates.

References

Elley, W.B. and Livingstone, I.D. (1972) *External Examinations and Internal Assessments*. Wellington: NZCER.

Hiscox, M.D. (1983) 'A Balance Sheet for Educational Item Banking'. Paper presented at the annual meeting of the NCME, Montreal. The quotation is from page 11.

Millman, J. and Arter, J.A. (1984) 'Issues in Item Banking'. *Journal of Educational Measurement*, 21:4, 315-330. The quotation is from page 320.

Prosser, F. (1974) 'Item Banking' in Lippey G. (Ed.) *Computer-Assisted Test Construction*. Englewood Cliffs, N.J.: Educational Technology Publications.

Wood, R. (1974) 'Question Banking' in Macintosh, H.G. (Ed.) *Techniques and Problems of Assessment*. London: Edward Arnold. The quotations are from pages 209 and 208.

Wood, R. and Skurnik, L.S. (1969) *Item Banking*. Slough: NFER.

Item Banks available from ACER

Australian Biology Test Item Bank

ACER 1984

Volume I: Year 11; Volume II: Year 12

Areas represented in the bank:

Volume I – Investigating the Living World

The Variety of Life, Organisms and Environments, Reproduction, Nutrition, Development and Growth, Populations, Interaction and Change in the Nautral World, The Living World.

Volume II – The Organism

Integration and Regulation of Multicellular Organisms, Cellular Processes, Heredity, Life – Its Continuity and Change, The Human Species, Science and the Scientific Process.

Items requiring the 'correct response' and the 'incorrect response' are represented in the Item Bank.

Australian Chemistry Test Item Bank

ACER 1982

Years 11 and 12

Areas represented in the bank:

Volume 1

Atomic Structure, Electronic Structure, The Periodic Table, The Mole and Chemical Formulae, Molecular Compounds, Infinite Arrays, Gases, Solutions, Surfaces, Stoichiometry, Head of Reaction, Chemical Equilibrium, Reaction Rates and Acids and Bases.

Volume 2

Redox Reactions, Electrochemical Cells, Electrolysis, Measurement and Chemical Techniques, Carbon Chemistry, Silicon Chemistry, Nitrogen Chemistry, Phosphorus Chemistry, Oxygen Chemistry, Sulfur Chemistry, Halogen Chemistry and Metals.

The *AIB: Mathematics Items* is currently being revised.

The *AIB: Social Studies Items* is currently out of print.

Item Banks available from the New Zealand Department of Education

Mathematics, Levels 1 to 9
Science, Forms 3 to 5
French, Forms 3 to 5
German, Form 3, Form 5.

Copying Permitted

© Copyright on this item is held by NZCER and ACER who grant to all people actively engaged in education the right to copy it in the interests of better teaching.

Combining Scores

By Alison Gilmore

All classroom teachers make quantitative assessments of how well their students are performing and frequently must combine marks from several different essays, tests, exercises or subjects to obtain an overall measure of achievement. At the simplest level, a teacher may combine the several marks for the essays or extended answers that make up a formal examination. At a second level, a single score may be required to summarise a pupil's performance over a year's study in one subject. For example, after a year of teaching science, a teacher may have end-of-term examination marks, and scores on assignments, practical exercises, laboratory reports and homework available. The teacher may also have measures of oral class participation and the like available for inclusion. At a third level, for the purpose of awarding certificates, for accrediting NZ University Entrance, for Queensland's Tertiary Entrance Score, for giving school prizes or scholarships, a student's overall assessment may be a combination of his marks in several different subjects. The way in which marks are combined may considerably influence the final assessment; in extreme cases it could make more difference than the way the students worked, or the way the teacher did the marking. In order to be fair to all students it is important to understand the factors which interact to affect the composite score.

The Validity of a Composite Score

When scores from different tasks are combined the specific information about how a student performed at a particular task is lost and the composite score provides only a *summary* of general performance. For example, the teacher may have Helen's marks for a number of tasks in French, such as, knowledge of grammar, conversational French, French literature, knowledge of French customs and way of life, and oral and written French. A composite score which condenses this information provides an indication of her overall achievement, but the actual absolute meaning of each of the scores has been obscured. She may be top in the class in her knowledge of French customs and way of life but extremely weak in her oral French. In day-to-day teaching, retaining separate assessments on the different tasks is often of more value than attempting to determine a composite score. The composite score masks a student's strengths and weaknesses and being in a summarised form it may be relatively meaningless. For example, Helen's overall mark in French which places her among the top thirty percent of students in her class disguises her extremely poor ability to speak French.

When the overall assessment is a matter of combining scores obtained on a number of distinct, but related tasks within a course of study, it is reasonable to assume that they are measuring attainment in the same area. This is the level two situation — like Helen's French. If you require a score which summarizes performance in a subject it is quite justifiable to combine marks because the composite score represents repeated measurements in the same discipline. It will in fact tend to be more reliable than a single mark. However, when the overall assessment is a matter of combining scores from *different* subjects the composite score that results has an even more limited meaning. The subjects may differ greatly in the demands they make on the students' knowledge and skills and in adding their marks together it is rather like 'adding four apples to six pears — this can only be done by calling them ten pieces of fruit and you no longer know what sort they are'.

There are occasions when teachers are required to provide a comprehensive rank order (order of merit) of students, for example, when accrediting New Zealand University Entrance, when awarding scholarships and when determining who is to get school prizes, these awards being based on a measure of overall academic achievement. This must be undertaken with considerable caution.

The essential feature to recognize in combining scores is that the measurement is essentially *relative*, not absolute. A composite score permits us to compare the standing of one individual against another, and to make judgements involving 'more' or 'less', but the real or absolute meaning of the scores is lost or masked.

Not All Scores May Be Combined

A teacher uses a variety of assessment procedures for a number of purposes. Not all 'bits' of assessment data should be considered as candidates for including in a composite score. Diagnostic tests, tests of mastery and

informal assessments of student progress are typically 'formative' and are useful as guides for further instruction; the assessment is on-going. These measures should *not* be combined., 'Summative' assessments, on the other hand, provide estimates of student achievement at the completion of a unit of work or at the end of term or end of year; they are less frequent and more comprehensive. Such assessments are norm-referenced and student attainment is considered in relation to that of a peer's. Norm-referenced assessments may be combined.

Determining a Composite Score

When two or more sets of scores are to be combined a decision must be made about the relative importance of each test, exam, assessment or task and its desired weighting in the composite score.

When you have made a decision about whether, say, French Vocabulary or knowledge of French customs is more important, then, if that decision is not reflected in the composite score you do not have a *valid* score.

Validity is a subjective judgement by the teacher, or group of teachers as to what weight each component shall be given. Firstly, judge the relative importance of each task. For example, if a teacher of English feels that the ability to speak well is more important than a knowledge of Wordsworth's poetry, a measure of a student's conversational skill should have greater weight than a score given for an essay on 'On Westminster Bridge'. Secondly, carefully examine the 'scope' of each component. If the score for a mid-year exam (assessing the first half of the year's work) is to be added to the score on an end-of-year exam (assessing the full year's work) what is the composite score actually representing? By simply adding the marks together, the first half of the year's work may receive greater weight in the composite than the second half of the year's work. It has been examined twice. Was the early course work more important than the later work? If it was not, the marks must be adjusted.

The importance or weighting of each component may need to be tempered by a consideration of the reliability of the various scores to be combined. Greater emphasis/weight should be given to more reliable measures. In general, reliability will be highest for a properly prepared objective test, moderate for carefully marked essays and lowest for informal, highly subjective appraisals of oral contributions and participation in class.

These considerations make a valid composite score possible and the next step is to make this possibility a reality.

Factors to Consider in Combining Scores

A test which is marked out of 100 will not count twice as much as (will not have double the weight of) a test marked out of 50. The marks when added together will simply weight themselves naturally but *not* necessarily in the desired way. Marking is essentially a relative (rank ordering) process and a quite different procedure for combining marks must be followed.

The essential features to consider when combining scores are illustrated in the following example.

Example:

Student	Test 1	Rank	Test 2	Rank	Total	Rank
Anne	50	5	40	1	90	1
Ben	53	4	35	2	88	2
Coleen	55	3	25	3	80	3
David	60	2	15	4	75	4
Eve	62	1	5	5	67	5
Maximum Possible Score	100		50		150	
Mean	56		24		80	
Standard Deviation	4.4		12.8		8.5	

Note: It is the teacher's *intention* that (in the composite score) Test 1 would count twice as much as Test 2.

It can be seen that the rank order of students in Test 2 is the reverse of that in Test 1, yet when the scores are totalled, the order for the composite score is the *same* as for Test 2 despite the teacher's intention that Test 1 should count twice as much. In other words, Test 1 has had no influence at all in deciding the final assessment (order) of the five students and the same result would have been obtained if only Test 2 scores had been used and the other marks ignored. Obviously such a situation is unsatisfactory and and needs to be corrected.

When two or more complete sets of scores are to be combined, the most important factor which influences the effect each will have on the final result is the *spread* (standard deviation) *of marks in each set*, not the possible maximum score, nor the mean (average) of the marks. The spread of scores in Test 2 (SD = 12.8) was approximately three times greater than that for Test 1 (SD = 4.4). Therefore, instead of the desired weighting of 2:1, the actual weighting was 1:3. That is, Test 2 scores had three times the influence in the composite score than Test 1. In general, then, the *more the marks are spread out, the greater will be their influence/weight in the composite.*

For a composite mark to reflect what the teacher intends, the spread of scores of the separate measures must be adjusted to reflect the appropriate relationships. Although a maximum possible score of 100 permits a greater spread than a maximum possible score of 50, it does not automatically follow that the sets of scores will weight themselves appropriately. In Example 1, for Test 1 although 100 was the maximum possible score the marks were tightly bunched around 55 while for Test 2, the scores were more dispersed.

Another factor to consider is the extent to which the various components are related (intercorrelated). In general, adjustments to the spread of a set of scores (to obtain appropriate weighting) is more important when the relationship between components is low or negative. For example, adjustments are very important when combining, say, marks for science and French, or scores on a test on valiency and a biology dissection. Adjustments are less vital when combining scores on tasks *within* a subject. French vocabulary and French Prose are more closely related, so departures in the natural weighting of components from their desired weighting is likely to be less serious than when scores from *different* subjects are to be combined. While it would be unlikely for a person who was top in one test in a given subject to be bottom in

another in the same subject area (as in Example 1 where the two sets of scores are perfectly negatively correlated) smaller discrepancies in rank will frequently occur and are to be expected given the measurement error present in even the most reliable of tests.

Procedure for Combining Scores

When all the students have taken the same series of tests, done the same assignments, or written on the same essay topics combining these marks is reasonably straightforward once a decision has been made about the weighting each score should have. However, when students do not all attempt common tasks, e.g., if they answer optional essays in an exam, or take different combinations of subjects, (as is usual in the senior secondary school), the measure of overall attainment must take into account the relative difficulty of each element as well as the variability of each set of scores. Is French 'harder' than Art? Is one optional assignment more difficult than another? And what happens if some marks are missing because Harriet was sick one day and Henry was moved into the class late in the year? These problems are taken-up in the following sections.

1. For cases when all students have done the same tests, assignments, etc., and done them all.

(i) *Estimate the spread of scores on each measure.* The simplest estimate would be the range (the difference between the highest and the lowest score), but, as this estimate is determined by two scores only, if just one student has done exceptionally well the range is quite misleading. The best estimate is the standard deviation. This index takes into account the spread of each score from the mean. Computing a standard deviation is a lengthy and tedious operation to do arithmetically. A calculator with statistical functions will make the computation easy. Without a calculator a good approximation to the standard deviation may be obtained in the following way:

1. Count the number of scores (the number of students who did the test).
2. Sum the top sixth of scores.
3. Sum the bottom sixth of scores.
4. Subtract the sum of the bottom sixth from the sum of the top sixth.
5. Divide this by half the number of scores (half the number of students).

In words: Estimate of Spread =

$$\frac{\textit{Sum of top sixth} - \textit{sum of bottom sixth}}{\textit{half number of students}}$$

(ii) *Determine the 'natural' weight of each set of marks,* that is, the ratio of their score spreads. For example, if the standard deviation for Test 1 is 4.4 and the standard deviation for Test 2 is 12.8, the natural weight of each component is 4.4 to 12.8, or approximately 1:3.

(iii) *Adjust each set of marks to obtain the desired weighting.* When each set of scores is to have *equal* weighting, their spread of scores should be approximately equal (the same SD for each). When the sets of marks are to have *different* weightings, their spread of scores must be in the same ratio as the weights required. If an examination mark is to be added to a term test mark and the examination is to count twice as much as the test, the ratio of the spreads of the two sets of scores would need to be 2:1 (e.g., examination SD = 4.4, term test SD = 2.2).

Adjusting the spread of a set of scores is simple: multiply or divide each score in a set by a constant. The spread of scores that results will be greater or less than the original spread by the factor that was used in multiplying or dividing. Thus, multiplying a set of scores by two will double the spread; dividing a set of scores by two will halve the spread, and so on. As all scores in a set are treated in the same way, their absolute value will change, but their relative standing or rank order will be unaltered. For this exercise, it does not matter if a test that was once marked out of 100 now gives scores greater than 100; the marks are not being treated as absolute but simply as an indication of which student did better than another on that task.

(iv) *Add the adjusted scores.* This gives you the rank order of students which represents a valid summary of student overall attainment. These composite scores may be converted to percentages, but it is important to remember that the percentage score has no more absolute meaning than the adjusted scores, but provides a more familiar set of figures to make comparisons between students.

2. When some data is missing, e.g., some students missed some tests, assignments, etc.

(v) *Add the adjusted scores (as in Step iv) and obtain the Composite Average.* Missing marks should *not* be treated as zero but may be handled by computing a *composite average* score: divide each student's composite total (the adjusted scores totalled, Step iv) by the number of components (tests, assignments, essays, etc.) for which there are scores.

A Worked Example: (see next page) The teacher wants to combine the scores on three tasks (an examination, a practical exercise and an assignment) so that the practical exercise and the assignment are weighed equally and the examination counts twice as much as the other two, 2:1:1.

The ratio of the 'natural' weights of the three elements is determined from their standard deviations: 4.0 to 8.0 to 4.2, that is, approximately, 1:2:1. The required ratio of weights, 2:1:1, may be most readily obtained by multiplying the examination scores by 2 and by dividing the set of scores for the practical exercise by 2. The scores for the independent assignment are unchanged. The three scores are then added together. To take account of absences the composite total is averaged to obtain one score which reflects the students' performance in the way you, as the teacher, have determined.

A comparison of the rank order when the unweighted marks are added together (J,N,I,L,K,M,G,H) and the

A Worked Example:

	Unweighted			Weighted				
Student	Exam	Practical	Assign-ment	Exam (×2)	Practical (÷ 2)	Assign-ment (unchanged)	Composite Total	Composite Average
Gerald	34	12	7	68	6	7	81	27
Helen	37	3	2	74	1.5	2	77.5	25.8
Ian	30	30	6	60	15	6	81	27
Jenny	42	20	—	84	10	—	94	47
Keith	—	16	12	—	8	12	20	10
Lynne	40	15	10	80	7.5	10	97.5	32.5
Mike	35	10	15	70	5	15	90	30
Noleen	32	25	13	64	12.5	13	89.5	29.8
Mean	35.7	16.4	9.3	71.4	8.2	9.3		
SD	4.0	8.0	4.2	7.9	4.0	4.2		

Note: Keith was absent for the exam and Jenny was absent for the assignment.

rank order when the scores are adjusted and appropriately weighted and when absences are allowed for (J,L,M,B,G,I,H,K) shows that all students except one have different places in class. This illustrates the importance of making adjustments to obtain a valid measure of overall achievement.

3. *For cases when students have done different combinations of subjects*

When students take optional tasks, such as, optional essay topics, or different selections of subjects, the ratio of score spreads alone (Step(iii) above) gives no guarantee that the effective weighting desired will result. The *average score* of each measure also becomes important. This is because differences in scores of students taking different components may reflect differences in the difficulty of the tasks, or differences in teacher marking standards, rather than simply differences in the ability of the students.

(vi) *Optional topics with equal weight.* One procedure for ensuring all components have equal weighting is to convert each set of scores to standard scores, that is, scale them to the same mean and standard deviation. This procedure has the effect of reducing each set of scores to a common scale which will have equal weighting when added together. Converting raw scores to standard scores may be done either by formula or by graph.

(i) *Standardizing scores by use of a formula.* The mean ($\bar{X}_s$) and standard deviation (SD_s) to which a set of raw scores may be scaled is flexible. A $\bar{X}_s$ of 60 and SD_s of 12 is frequently suggested as suitable for assessment programmes in the school. This distribution provides scaled scores normally within the limits of 0 and 100. Tables which convert raw scores to this scale have been published (e.g., Queensland Department of Education, 1972, pp.62–65) and score conversion is therefore reasonably straightforward. T-scores with a $\bar{X}_s$ of 50 and SD_s of 10, are also frequently used as an alternative scaled score distribution.

Example: To compute the scaled score for each student the following information is necessary.

1. The student's raw scores (X_R)
2. The mean of the set of raw scores ($\bar{X}_R$)
3. The standard deviation (or estimate) for the set of raw scores (SD_R)

The scaled score is then computed as:

Step 1. Subtract the mean from the raw score ($X_R - \bar{X}_R$)
2. Divide the difference by SD_R
3. Multiply the result of (2) by the scaled SD_s
4. Add the result of (3) to the scaled mean ($\bar{X}_s$)

That is:
$$X_s = \bar{X}_s + SD_s \frac{(X_R - \bar{X}_R)}{SD_R}$$

If $\bar{X}_s$ and SD_s are to be 60 and 12, respectively, then, for example, if the mean and standard deviation for a set of raw scores is $\bar{X} = 25$ and $SD = 5$, a raw score of 35 is converted to a scaled score as:

$$X_s = 60 + 12\,\frac{(35 - 25)}{5} = 60 + 12\,(2) = 84$$

and a raw score of 18 is converted to a scaled score as:

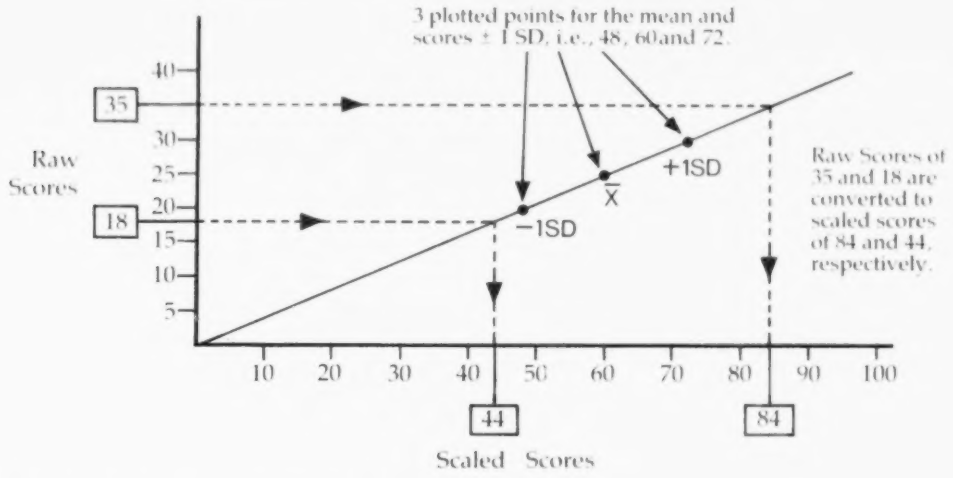
$$X_s = 60 + 12\,\frac{(18 - 25)}{5} = 60 + 12\,(-7/5) = 60 - 16.8 = 43.2$$

If a conversion table, such as the one referred to above is available, the only computation required is that in step 1. The difference score obtained is then entered into the appropriate table and the scaled score read off.

(ii) *Standardizing scores by use of a graph.* A graph is drawn with one axis representing the raw scores and the other representing the scaled scores. Such a graph can be used at a variety of levels of sophistication. In its simplest form three score equivalents are selected to represent the mean and scores one standard deviation above and one standard deviation below the mean for the set of raw scores and the corresponding scaled scores. In order for all the raw scores to be scaled, the three points which correspond to these three pairs of scores are plotted. Next, a

line is drawn through the three points; the graph can then be used to convert any given raw score in a set of scores to a scaled score.

Example: A set of raw scores, with a mean of 25 and a SD of 5 is to be converted to a set of scaled scores with a mean of 60 and SD of 12.



Again, the absolute value of scores will alter but the relative position (rank) of each student remains unchanged.

A Worked Example: Essays 1 and 2 are optional. Essay 3 is compulsory. Each essay is to have the same weight.

Student	Raw Scores Essay 1	Essay 2	Essay 3	Scaled Scores Essay 1	Essay 2	Essay 3	Composite Total	Composite Average
Alan	28		35	46		35	81	40.5
Betty	36		30	62		30	92	46
Cathy	32		24	54		24	78	39
Dianne	24		26	38		26	64	32
Ellen		21	18		62	18	80	40
Fred		9	15		38	15	53	26.5
Gwen		17	12		54	12	66	33
Harry		13	7		46	7	53	26.5
Mean	30	15.0	20.9	50	50	20.9		
SD	4.5	4.5	8.9	8.9	8.9	8.9		

For all essays to be equally weighted in the composite score, each set of scores may be converted to a common distribution (e.g., $\bar{X}_s = 60$, $SD_s = 12$) and then added. However, because Essays 1 and 2 are optional components and Essay 3 is common to all students, all that is necessary for equal weighting is (a) that the spread of scores for the three essays be approximately equivalent and (b) that the mean scores for Essays 1 and 2 be equivalent. This may be achieved by scaling the scores for Essays 1 and 2 to, e.g., a $\bar{X}_s$ of 50 and a SD_s of 9.0 (the same SD for Essay 3). The adjusted scores are then added together and averaged to obtain the composite average score. Thus: *When components are to have equal weights but not all students do the same tests, essays, subjects, etc., any particular standing/rank must have the same numerical value in the optional tasks.*

(vii) *Optional components with different weightings. Combining scores from different subject areas.* It is argued (e.g., Thyne, 1974) that by their very nature, optional tasks must be considered of equivalent importance. This may be the case for optional tasks within a course of study in which the tasks are judged to be of comparable difficulty. However, the same assumption of comparable difficulty cannot be extended to the situation in which scores from different subjects are to be combined. While there is extensive debate about the practice of combining scores from different subjects, the complexity of the problem increases when not all students study the same subjects but have different combinations of subjects as is common in the senior secondary school.

The typical situation which arises is that students are to be compared with each other for some scholarship, award or certificate and yet they have taken different combinations of subjects. We know that

different subjects demand different abilities and so some attempt to take account of 'quality' differences in the students is essential if scores in different subjects are to be added. In order to establish comparability of abilities between different subject groups some form of moderation is required.

The essential feature of moderation is to determine the relative level and spread of abilities of students in different groups. This may be accomplished in a number of ways. An *appropriate* reference test* may be administered to all students. The raw scores for each group of students may then be scaled to the mean and standard deviation of scores that group obtained on the reference test. As before, the mean and scores one standard deviation above and below the mean on each measure may be used as points for constructing the graph from which the raw scores may be scaled.

In the absence of a common reference test, an alternative procedure for establishing comparability of abilities is to scrutinize groups of students who are doing the same combination of subjects and check that the mean scores are the same. If, for example, the mean score for students studying German is 50 but the mean score for geography, English and economics for the same group of students is 39, 40 and 41, respectively, the German marks may be scaled to a mean of approximately 40.

These approaches have some limitations, particularly for those subjects which may rely heavily on special abilities not common to other subjects, e.g., music, technical drawing and art. The students may be very able in that particular discipline without being equally able in other academic subjects.

A Worked example: The following estimates of overall achievement in each of five subjects are listed (i). An 'appropriate' reference test has been administered to all students. The mean and standard deviation ($\bar{X}$, SD) for each group of students for French, Maths, history, physics and German are (55, 10), (55, 15), (50, 20), (65, 5) and (60, 10), respectively.

All scores in each subject are scaled to the appropriate mean and standard deviation and then summed. Because the students are presenting different numbers of scores for combination, an *average* composite score is then computed. In this way, decisions about students' overall attainment in relation to that of his peers may be made fairly and validly.

A Worked Example:

Student	French	Maths	Hist	Physics	German	Scaled scores French	Maths	Hist	Physics	German	Composite Total	Composite Average
Alan	70		65		40	68		56		46	170	56.7
Betty		24		42	45		38.5		63.4	53	154.9	51.6
Cathy	60		70			61		64			125	62.5
Dianne	30	78		70		42	77.5		72.5		192	64
Ellen			75		60			72		74	146	73
Fred		32		30			44.5		59		103.5	51.8
Gwen	40		60		50	49		46		60	155	51.7
Harry		50	40	50	55		58	14	65.5	67	204.5	51.1
Mean	50	46	62	48	50	55	55	50	65	60		
SD	15.8	20.7	12.1	14.6	7.1	10	15	20	5	10		

This paper has attempted to alert the reader to the fundamental issues and procedural steps in the combination of scores so that a fair and valid estimate of overall attainment for the individual student may be achieved.

*Extensive debate surrounds the issue of appropriate reference tests for moderating purposes. It is not proposed to enter into a discussion of the variety of measures possible. The reader is referred to Elley and Livingstone (1972) for a treatment of this topic.

Notes

Alison Gilmore is a Test Development Officer with the New Zealand Council for Educational Research.

The quotation about apples and pears is from the Schools Council (UK) Examination Bulletin 32, mentioned below.

References and Suggested Reading

Department of Education, Queensland, *Moderation Within Schools*, Research and Curriculum Branch, 1972.

Dunn S.S. *Measurement and Evaluation in the Secondary School*, ACER, 1967.

Elley W.B. and Livingstone I.D. *External Examinations and Internal Assessment*, NZCER, Wellington, 1972.

Glock M. and Millman J. "The Assignment of School Marks" in Granlund N.E. (Ed) *Readings in Measurement and Education*, The Macmillan Company, London, 1968.

Payne D.A. *The Assessment of Learning*, D.C. Heath and Company, Massachusetts, 1974.

Peddle B. and White G. *Testing in Practice*, Heinemann Educational Books, Auckland, 1972.

Scannell D.P. and Tracy D.B. *Testing and Measurement in the Classroom*, Houghton Mifflin Co., Boston, 1975.

Schools Council, *Assessment and Testing in the Secondary School*, Schools Council Examination Bulletin 32, Evans/Methuen Educational, London, 1975.

Thyne J.M. *Principles of Examining*, University of London Press Ltd, London, 1974.

EVALUATING WRITING

A

New GUIDE

TO THE

English Tongue:

In Five PARTS.

CONTAINING,

I. Words, both *common* and *proper*, from *one* to *ſix Syllables*: The ſeveral ſorts of *Monoſyllables* in the *common* Words being diſtinguiſhed by Tables, into Words of *two*, *three*, and *four* Letters, *&c.* with ſix ſhort Leſſons at the End of each Table, not exceeding the order of Syllables in the foregoing Tables. The ſeveral ſorts of *Poly ſyllables* alſo, being ranged in proper Tables, have their Syllables divided, and Directions placed at the Head of each Table for the *Accent*, to prevent *falſe Pronunciation*; together with the like Number of Leſſons on the foregoing Tables, placed at the End of each Table, as far as to Words of *four* Syllables, for the eaſier and more ſpeedy Way of teaching Children to Read.

II. A large and uſeful Table of Words, that are the ſame in *Sound*, but different in *Signification*; very neceſſary to prevent the Writing one word for another of the ſame *Sound*.

III. A ſhort, but comprehenſive *Grammar* of the *Engliſh* Tongue, delivered in the moſt familiar and inſtructive Method of *Queſtion* and *Anſwer*; neceſſary for all ſuch Perſons as have the Advantage only of an *Engliſh* Education.

IV. An uſeful Collection of *Sentences* in *Proſe* and *Verſe*, *Divine*, *Moral*, and *Hiſtorical*; together with a ſelect Number of *Fables*, adorn'd with proper Sculptures, for the better Improvement of the Young Beginner. And

V. *Forms* of *Prayer* for Children on ſeveral Occaſions.

The *Whole*, being recommended by ſeveral *Clergymen* and eminent *Schoolmaſters*, as the moſt uſeful *Performance* for the Inſtruction of *Youth*, is deſigned for the Uſe of SCHOOLS in *Great Britain* and *Ireland*.

The EIGHTH EDITION.

By THOMAS DILWORTH,

AUTHOR of the

SCHOOLMASTERS ASSISTANT; and *Schoolmaſter* in *Wapping*.

PHILADELPHIA:

Printed and Sold by B. FRANKLIN, MDCCXLVII.

Evaluating Writing

By David Philips
NZCER

Introduction

Teachers need marking techniques. Plenty are available, but which are the best? That depends on what you want them for. If you want to assess

1 Normal coursework writing — what the children do every day — then there will be two jobs for the marking to do:
   (a) diagnose faults (so that we can give exercises to correct them)
   (b) ascertain progress (so we can see if our teaching is successful).

If you want to assess

2 A year's work — or even a term's — then we will be looking for a technique which will:
   (c) assess the child's progress compared with his or her earlier performance
   (d) possibly provide a comparison of performance against the rest of the class, or the rest of his or her age group.

Choose your assessment technique carefully — it must fit the task (one of the above four), the class level, and the pupil. 'In evaluating writing we are assessing much more than their grasp of a programme: we are evaluating the students themselves.'

Why is Writing Difficult to Assess?

Despite the excellent research of Janet Emig, Donald Graves and others, the writing process itself is still largely a mystery. We know that it is a very complex process requiring the mastery of a variety of interrelated skills. Apart from the essential inputs of reading and thinking, skills such as knowing how to organize material, awareness of the teacher's goal, understanding the purpose of the specific writing task, all play an important part in creating written material. It is not surprising therefore, to find that pupils vary considerably in their ability to write. While some pupils improve their writing with relative ease, others consistently find writing a difficult enterpris.. It is important to remember, though, that writing skill *develops*. It is not a static ability which one either has or has not. Consequently the end-point reached will inevitably vary from one person to the next. Since writing skills are usually in a state of change, and fluency takes time to develop, it is essential that both diagnostic and end-of-the-year assessments be made with the intention of encouraging the burgeoning writer.

Writing has both 'deep' and 'surface' features. The 'deep' ones include the purpose of the writing, its content and structure. The 'surface' ones are the orthographic or transcriptional aspects of spelling, punctuation, capitalization and grammar. It is so easy for teachers to focus on the 'surface' features and so easy for the pupils to think they are the only important aspects that both teachers and pupils may lose sight of the basic purpose of the writing. Collins and Gentner have drawn attention to this phenomenon, and have labelled it 'downsliding'.

> One great difficulty for writers is maintaining *connective flow*. The relationships between ideas must be made clear. Yet in order to write about an idea, the idea must be expanded downward into paragraphs, sentences, words and letters. Sometimes writers — particularly children — become lost in the process of downward expansion and lose sight of the high-level relationships they originally wanted to express. *Downsliding* — the phenomenon of getting pulled into lower and more local levels of task processing — is a very common problem in writing and in other domains as well. If a teacher emphasizes accuracy in spelling and grammar it will reinforce the natural tendency toward downsliding. The overall result will be that children focus almost exclusively on lower-level task components when they write.

Of course, it is often very difficult to avoid emphasizing those features of pupils' writing which are most clearly in error. But it would be unfortunate indeed if the error-seeking red pen was not tempered with a sympathetic attempt to improve writing skills beyond the merely 'surface' characteristics. It is not an easy job to mark the 'deep' features. But they do have to be assessed if we are to be helpful.

1 Evaluating Performance During the Year

Assessing the Developing Writer

Writers differ in their learning rate and in their potential for improvement. However, there is little point in prejudging a pupil's likely achievement in writing and teaching to that expectation. Instead, try to pay close attention to overall development and focus on specific writing difficulties. 'Composing a piece in any mode is a complex linguistic, experimental, cognitive, affective and scribal act.' (Cooper, 'Measuring Growth in Writing', *English Journal*, Vol. 64, No. 3, March 1975, p. 112.) Ask yourself:

Surface problems

Has this pupil an adequately legible style of handwriting?
How extensive is the pupil's command of language? Is she or he having difficulty with spelling, subject-verb agreement, sentence structure? Has she or he had sufficient practice with this mode of writing?

Deep problems

Can this pupil stand back from present circumstances and order thoughts in an appropriate manner? Does he or she know how to compose written work? Is choice of content (within the piece) or organisation of the content giving problems?

Has the pupil had enough experience to write on this topic? Is the pupil sufficiently motivated to write on this topic? Is he or she having difficulties with parents, peers, etc., which might affect performance?

This preliminary look may reveal that the student has difficulties. If so, steps will have to be taken to provide appropriate assistance. The teacher using this technique interprets the pupil's writing as part of a complex series of interrelating factors, each an integral part of his writing ability. Further, the pupil's progress is gauged against several variables. A mark in the teacher's markbook which is a simple sum of the number of errors the pupil has made, is not nearly as useful.

Revision is an integral part of most writing. Therefore another important procedure to follow is to allow the student to revise and re-work part of the writing if necessary, and consult peers, and the teacher, on the content and form of the written work. (The research of Donald Graves in this area is especially instructive. Though it deals with pupils in their first few years at school the conclusions are universal.) This procedure allows pupils to view writing as a continuous process with several mutually supportive stages, rather than simply as a one-off type of exercise done merely for the teacher's benefit.

Assessing the student's work during the year will entail these activities, in this approximate order:

(1) Consider the surface and deep features on page 2. Carefully note the pupil's development or behaviour within each area.

(2) If the content of the pupil's writing seems unrelated to the topics given, check the questions you asked, and the instructions you gave, for ambiguity. Make sure that the tasks set are within the students' capabilities, yet challenging.

(3) Discuss the more immediate difficulties with the student; provide a willing ear; be supportive.

(4) Correct 'surface' errors but by focussing on only one or two specific examples each time until the student reaches an appropriate level of mastery in them, e.g., capital letters for a few days or weeks, then commas.

(5) Take remedial action where necessary over specific thorny problems (i.e., by giving extra instruction and help, for example, with persistent poor spelling).

(6) Keep a careful written record of the student's improvement in addition to the first 'diagnosis'; update it regularly (e.g., 3 or 4 times a term).

Although this strategy requires considerable care, it is designed to encourage the student in a positive fashion rather than to inhibit development. Comments on the pupils' writing, whether verbal or written, should be selective rather than comprehensive. This is so that the pupil can focus on separate aspects of performance and gradually bring about improvement in them.

Methods of Marking

If marks or grades have to be awarded to pupils' written work, bear in mind some of the findings from research on the marking of essays. Even though most of this research has been concerned with secondary or tertiary level students, it is a useful reminder of the fallibility of the most carefully prepared teacher!

It has been established, for example, that the same piece of written work will not always receive the same mark, even when marked by the same marker. The order in which several essays are assessed may affect the quality of the mark awarded; thus a series of good essays may build up the marker's expectations so that when a poor essay comes along it will obtain a much lower mark than if it had been preceded by a series of mediocre essays; the reverse also applies. If essays are marked over several days, by the last day of marking the assessments are likely to be much less consistent than they would have been earlier in the piece. However, this is unlikely to be a serious problem when marking occurs in a single session, and provided class papers are not always marked in the same order. If papers are marked in the same order (e.g., alphabetically or by designated groups), the biases introduced due to marking order are likely to be significant.

A more pressing problem for the classroom teacher is deciding which criteria ought to be applied to any given piece of writing. What features should be examined? How inadequate does a pupil's performance have to be before some kind of assistance becomes necessary?

(i) Revealing Criteria of Marking

Complete agreement on the most appropriate features to assess does not exist. Different markers give more or less weighting to different criteria. For example, two secondary school English teachers may each have a pupil who insists on using an ampersand (&) instead of writing 'and' in his essays. The first teacher may consider this abbreviated technique to be a major breach of convention, and mark the pupil more harshly as a result. The second teacher may well ignore the ampersands and when handing back the pupil's essay simply make a passing reference to it. Some markers are consistently bothered by spelling mistakes: the attitude seems to be that incorrect spelling has to be stamped out, so the red marks will fly onto the pupil's essay. Although these examples may appear to be relatively trivial, research has shown that the consistent breaking of the conventions of spelling and punctuation can lead to reduced marks since the number of errors (even though they might be minor ones) inhibits the marker and also directs his or her attention away from the quality of ideas or content of the writing. Many studies, for example, have shown that handwriting quality also has an influence on the marks awarded to essays so care is required to ensure that students with poorer handwriting, spelling and punctuation do not suffer in their marks as a result.

Another problem, and one pupils often bewail, is marking criteria being inconsistently applied. Naturally, teachers apply different criteria depending on the aims of a particular writing exercise. A piece of creative writing such as a short story is likely to be examined for its quality of ideas, since any writing inaccuracies can always be tidied up. After all, published writers have the service of editors and secretaries. On the other hand, a piece of descriptive writing (e.g., an account of a holiday, or the construction of a familiar object) is more likely to be assessed on the basis of the accuracy of the

events recounted or the orderly discussion of the steps involved in the activity concerned. At the secondary level, essays may well be examined for their structural features: how well ideas hang together, whether the topic is appropriately introduced and covered to an adequate extent, etc. For all these types of exercise, the presentation (legibility, appropriate location of headings and margins, etc.) and the orthographic features (spelling, punctuation, grammatical accuracy) while part of the 'total communication', are *not* the most significant elements in the overall pattern of writing development. In any assessment scheme, therefore, they should not assume undue importance.

To sum up, the first step is to clarify the *purpose* of the piece of writing which is to be assessed. Some common purposes (following Stibbs) are: the recording of information for the writer's own use; recording information for someone else's use; helping the writer to sort out his own experience and thoughts; helping the writer to understand the experiences of others; symbolising experience in particular ways; describing; instructing; persuading.

The writing itself may be in any of several *forms* (such as notes, summaries, reports, poems, plays, stories, descriptive accounts of people, places or objects, letters or lists of instructions), so the criteria of assessment will need to be adapted to suit both the form of the writing and its purpose. A set of instructions, for example, would need to be well laid-out and sequenced accurately for ease of interpretation. Assessment would, therefore, tend to emphasize those features. On the other hand, an essay about a recently read book might be assessed according to how well the writer summarises the book's contents and dicusses his or her own reaction to it. Paragraphing and coherence would also be important.

Teachers must ensure that their pupils *know* what is going to be examined in their written work: for example, that this is a descriptive piece and accuracy of information and orderly discussion will count highest. Although it is often said that writing is a game and a test of one's ability to guess what the teacher wants, this attitude is not a worthy one. Criteria should be made explicit, and a careful watch has to be kept to make sure that unconscious criteria are not assuming greater importance than stated ones. To this end, markers need to (a) expose their 'standards', through self-examination; (b) communicate their criteria to their pupils so that the pupils can take them in; (c) keep a careful record of the *kinds* of comments they make on each pupil's 'essays' and of what they have done to assist the pupil's improvement.

(ii) Features of Writing

Some elementary distinctions are useful.

Mechanics

The 'surface' features mentioned before are often known as writing mechanics, or transcriptional features, since they represent those aspects of writing which are readily recognized as the basics of written communication. They include:

a. *Handwriting*
The legibility of the writing will range from uninterpretable to absolutely clear and easy to read. As it is usually the first feature of a piece of writing to be noticed (except, perhaps, for the overall layout of the whole communication), and creates an impression in the reader's mind about the writer's attitude to his or her task, it is easy to be misled by it. Unless the pupil is being assessed on handwriting alone, there seems to be little justification in making it part of any evaluation of writing quality, however hard the temptation to do so might be.

b. *Punctuation*
Inappropriate punctuation (ranging from the occasionally omitted comma to inability to distinguish one sentence from another — Mina Shaughnessy provides some excellent examples of such problems in *Errors and Expectations*) is another immediately recognizable feature of pupil's writing, found as much in university students' writing, it seems, as in primary schools. From the marker's point of view, continually misplaced commas and/or full-stops are a jarring note in any writing (with the exception of deliberate experiments with language as in some forms of 'creative' writing), since they actively impede comprehension.

c. *Spelling*
Incorrect spelling is another easily identifiable feature of writing, which many markers include as part of their assessment. The range of performance will be from no spelling mistakes to a plethora of errors. As with illegible handwriting, spelling mistakes give markers a hard job as they tend to counteract any positive impressions they might hold about a piece of writing.

d. *Grammatical Usage*
Wrong tense, wrong pronoun, inappropriate subject-verb agreement or other incorrect forms of words can also be labelled 'surface' features since they are easily identified and frequently commented upon, but seldom have the effect of destroying ideas or logical sequence.

e. *Sentence Structure*
This element is often counted as a 'surface' feature, including such things as sentence fragments, over use of 'and', misrelated clauses, etc. However, many of these aspects can be interpreted as punctuation difficulties or awkward usage.

While these features can easily impede understanding, and are often referred to as carelessness, they have very little to do with the *content* of a piece of writing, unless together they so obscure a writer's message that it cannot be understood at all, or only with extreme difficulty. It is best not to assess the quality of a piece of writing on this basis alone.

Content

The 'deep' features, however, are much more difficult to assess, and it is at this point that markers begin to diverge even more widely. Any balanced assessment needs to include a careful appraisal of these aspects. The problem is not so much that markers disagree about the choice of criteria but that they attach different weights to different traits. Although this is virtually an insoluble problem the most significant 'deep' features which ought to be considered in any assessment of writing are listed below without any attempt at ranking their importance in relation to each other.

a. *Ideas*

This feature includes qualities such as relevance, accuracy, fullness of treatment and originality of approach. However, it is often extremely difficult to assess the adequacy of a pupil's treatment of a topic. The negative features are often as prominent as the positive: irrelevant ideas, inaccurate representations of facts, excessive emphasis on insignificant points, a confused attitude towards the topic, etc. On the positive side satisfactory responses often differ a great deal in their treatment of the topic; how easy it is to give high marks to an essay in which the point of view agrees with your own and to penalise different approaches! It is also important to strike a balance between sheer volume of ideas and the quality of the ideas — hence the importance for some markers of the rather nebulous feature called originality.

b. *Organization*

A survey conducted by the author in 1979 revealed that university essay markers considered organization of material to be the biggest stumbling-block for many writers. The development of the ideas: how they are structured within the essay, appropriately dividing ideas into paragraphs, using contrast, introducing the main features of the topic, putting ideas in an appropriate order are all part of this feature. The haphazard grouping of ideas is likely to be assessed somewhat harshly by many teachers, while writing which 'flows' will probably be given higher marks. Markers should take care to be consistent in assessing this feature and consider if 'flowing' is more important than having new and powerful ideas.

c. *Word Choice*

Aspects of this feature are the use of appropriate terminology (i.e., adapted to the presumed audience of the writing); words which can be readily understood, with definitions included when deemed to be necessary; the avoidance of ambiguity, hackneyed expressions and redundancy; and the use of concise, clear words rather than long, obscure ones. Marks must depend to a certain extent on the clarity with which the purpose or context of the writing was made clear to the students.

d. *Style*

Perhaps the most difficult feature to assess is the 'flavour' of a piece of writing, i.e., how well the writer sustains his attitude or commitment, the suitability of the writing for its intended purpose and audience, the use of stylistic devices and the fluency displayed. Judgements on style are most likely to be highly subjective. The range of possibilities confronting the writer is very wide, and the effects of style on the marker are subject to influences beyond knowing.

The extent to which these features play a part in the overall assessment of the quality of a piece of writing remains a matter for individual teachers to determine. It is worth bearing in mind, however, that even though elaborate marking schemes (some of which are discussed in the following section) have been developed, the problem of whether a particular piece of writing *meets* the criteria or not still exists.

(iii) Marking Schemes

One of the hardest tasks an English teacher faces is deciding which aspects of writing are most important. For example, is style most important, or are the ideas the writer is putting forward more so? Some of the marking schemes currently in use will be briefly covered in this section in order to assist thinking about this problem.

Broadly speaking, there are two types of marking schemes, holistic (or impressionistic) and analytic (or atomistic). In analytic marking, a series of judgements is made about the pupil's writing according to a set of clearly specified criteria. Marks are awarded for each criterion or essay feature according to a predetermined scale, up to a stated maximum. This is probably the most useful approach for evaluating work done during the year, when diagnosis and appropriate assistance are most important. Impressionistic marking, on the other hand, simply requires a single judgement about the quality of a piece of writing, and is most useful for end-of-year assessments (see later section on Holistic Marking).

Analytic Marking

As an example of an analytic marking scheme, take a recent project undertaken in Canada, which developed criteria for the evaluation of different modes of writing for grades (years) 7 and 8 (Forms 2 and 3). Each criterion has been elaborated to make it easy to divide work into the categories of high, medium and low. The introduction includes the comment that 'we should like to see both teachers and students sensitive to the fact that certain writing tasks call for different styles, different language choices, and attention to particular skills each related to the *function* or purpose of the writing and the intended *audience*'. To illustrate the criteria, here is an excerpt from *Word Choice:*

Imaginative and Varied Language Choices: Grade 8

High:	Words and images which provide sharp and concrete pictures for the reader are frequent. Occasional experiments in stretching vocabulary and images to include new or unusual words or images. Trite expressions are usually eliminated. Flowery excesses — too many adjectives/adverbs piled on top of each other — are avoided.
Medium:	Generally word and image choice is at a more ordinary level with some experimentation, not always successful, in vocabulary expansion or creation of an image. The student still lacks full control and some excesses or redundancy may occur as well as the occasional trite, hackneyed expression.
Low:	Little experimentation with language. Reliance on the trite and very ordinary bland or abstract expression. Occasional errors in the use of standard vocabulary.

This publication includes the criteria Organization, Word Choice, Conventions and Mechanics, Content/Ideas and Style, and also includes criteria related to specific modes in writing such as Narrative: Eye-witness account, real or imagined; Narrative: Second Person, with emphasis on description; Narrative: Third Person, emphasis on dialogue; and Exposition: Presentation of a viewpoint or argument (which covers six qualities — planning, argument, style,

sentence style, fairness or objectivity and freshness/originality). However, no criteria are suggested for 'free' writing, book reviews, reports, etc. It is also suggested that a scoring scale could be used, with pupils receiving points for each criterion as follows:

Organization	2	4	6	8	10
Language Choice	2	4	6	8	10
Sentence Variety	1	2	3	4	5
Grammar	1	2	3	4	5
Spelling	1	2	3	4	5

Possible score range: 7-35, if a composite score is thought useful.

One of the most well-known analytic scales is that of Diederich, as discussed in *Measuring Growth in English*, which looks like this:

	Low		Middle		High	
General Merit						
Ideas	2	4	6	8	10	
Organization	2	4	6	8	10	
Wording	1	2	3	4	5	
Flavour	1	2	3	4	5	____
Mechanics						
Usage	1	2	3	4	5	
Punctuation	1	2	3	4	5	
Spelling	1	2	3	4	5	
Handwriting	1	2	3	4	5	____
					Total:	____

In addition to the table of points, a general description of high, medium and low performance is given for each criterion. Under 'Organization', for example, is this description:

High: The paper starts at a good point, has a sense of movement, gets somewhere and then stops. The paper has an underlying plan that the reader can follow; he is never in doubt as to where he is or where he is going. Sometimes there is a little twist near the end that makes the paper come out in a way that the reader does not expect, but it seems quite logical. Main points are treated at greatest length or with greatest emphasis, others in proportion to their importance.

Middle: The organization of this paper is standard and conventional. There is usually a one-paragraph introduction, three main points each treated in one paragraph, and a conclusion that often seems tacked on or forced. Some trivial points are treated in greater detail than important points, and there is usually some dead wood that might better be cut out.

Low: This paper starts anywhere and never gets anywhere. The main points are not clearly separated from one another, and they come in a random order — as though the student had not given any thought to what he intended to say before he started to write. The paper seems to start in one direction, then another, then another, until the reader is lost.

As an example of a 'surface' characteristic, the descriptions for 'Handwriting Neatness' are as follows:

High: The handwriting is clear, attractive, and well spaced, and the rules of manuscript form have been observed.

Middle: The handwriting is average in legibility and attractiveness. There may be a few violations of rules for manuscript form if there is evidence of some care for the appearance of the page.

Low: The paper is sloppy in appearance and difficult to read. It may be excellent in other respects and still get a low rating on this quality.

What these and similar 'analytic schemes' share is a reasonably elaborate description of those essay features expected for levels of writing quality. Although a composite score can be obtained for any piece of work analysed in this way, it is not likely to be very useful since pupils with the same mark could vary greatly in their handling of the individual features. Analytic marking, therefore, is most useful in classroom assessment when the reasons for the separate marks awarded are clearly explained to the pupil. If the application of this technique revealed class-wide deficiencies in one or other skill areas, further teaching could be organized to cover these points, as a back-up to the informal teacher-student dialogue conducted throughout the year.

A single mark or grade made by amalgamating all the analytic scores, however, is an insufficient indication to a pupil of his writing progress. Written comments would have to be added as well, in which a careful evaluation was made of both the good and inadequate aspects of the pupil's performance on that task. Diederich, for example, has shown that *the procedure with the most consistently positive effect on students' motivation is to correct* **one** *particular type of error, and to provide a comment on* **one** *particular strength in the student's piece of writing.* In this way the comments are more likely to be taken to heart and kept in mind by the student, particularly if they are presented in an encouraging manner. A study conducted by Page showed that students who receive individualized comments from the teacher obtain the highest scores, compared to students receiving automatic, impersonal comments (e.g., 'Good Work') or only a mark. The relationship between supportive feedback and student improvement is a subtle one, and Diederich's advice is especially worth noting.

2 Evaluating the Year's Performance

End-of-year grades or marks are not an integral part of the learning process. But they do provide an *estimate* of the amount and kind of learning achieved by the student, as their main function is usually to distinguish students from each other, to provide a comparison or ranking.

Holistic Marking

Evaluating writing skills is difficult because of the integrated nature of a piece of writing and differences in markers' approaches. The assessment technique which takes this into account is impressionistic (or holistic) marking. Research has shown that a rapid overall judgement of the quality of a piece of writing is as *reliable* a technique as the much slower method of analytic marking. Using this holistic technique, the marker reads quickly through each pupil's script in order to assign a mark or grade to it on the basis of his or her view of an adequate performance. Separate assessments of individual features are not made.

As a check on the consistency of the marking, essays can be sorted into three approximately equal piles representing good, average and poor efforts, with each of these piles being sorted again into three piles, making nine in all. Thus essays in pile 3 can be compared with pile 4, etc., to ensure that (a) there are differences in quality between each of the neighbouring piles and (b) essays within each pile are of similar quality. With practice this checking process can also be completed relatively quickly.

When a team of markers is involved in this activity, checks are required to ensure that all the markers have comparable standards. Normally this is done twice: once before any assessments are made so that everyone involved knows what is being looked for (that is, the criteria of adequate performance) and, secondly, after the assessments have been made in order to check for any large inter-marker differences. The range of marks awarded by each marker needs to be examined too. Obviously, some markers are more harsh in their judgements than others, and may use a more restricted range of marks in which, for example, the high ones tend to be avoided except perhaps for an outstanding response. Others will be more lenient, and may fail only students with excessively poor answers. Some bunch their marks around the middle. Consequently, it is necessary to be very clear about the characteristics expected of answers at each point of a scale, and to ensure that each marker agrees with them prior to assessment. Even then differences will probably occur. But although personal biases can never be completely removed, working closely with other teachers will assist the process of ironing out both foreseeable difficulties and any systematic bias due to identifiable idiosyncrasies.

What other sources of variation can be guarded against? The questions students are required to answer need to be devised very carefully. Rosen, for example, has shown that in a list of essays, from which a pupil is required to choose only one, different essays may make very different linguistic, content and organizational demands. It has also been shown that students, when given a choice of questions, do not necessarily answer the ones they can obtain their best marks on. Ambiguity in question phrasing has to be guarded against, too, as some pupils may interpret their tasks quite differently from other pupils when confronted with the same essay question, and do badly.

With especially important examinations, it is sometimes a healthy practice to use more than one marker. This reduces personal bias and, where a pupil has interpreted a question in an unusual fashion, for example, provides an alternative opinion of the quality of the pupil's writing. Multiple marking of the same papers is generally preferable to a single rating and does not take a long time when the impressionistic technique is used. It also results in greater consistency between markers in their assessments.

3 Performing an Evaluation: A Checklist

Consider these points carefully:

I *Why* are you making the evaluation?
Remember that initial assessments serve a different function from those made during the year, and especially from those which attempt to sum up a whole year's work. For example, is your evaluation designed to provide an overall judgement of a pupil's writing ability? If so, ideally it will be based on a *range* of writing tasks, as one task alone is hardly representative.

II What do you hope the *outcome* will be?

The way the information obtained will be used is probably more important than the method adopted. Is it mainly to help your students improve their writing skills, to widen your knowledge of their abilities, or to provide a means for comparing students with each other?

III Choosing appropriate techniques:

a. To obtain a deeper understanding of your pupils' writing ability, ask yourself the questions listed on page 2.
b. To assist pupils to improve their writing, follow the procedures listed on page 3.
c. When a mark is required on a piece of writing done during the year, work carefully from a set of explicit criteria. The features listed on pp. 4–5 will assist here, though they will have to be adapted for different class levels. The marking schemes on page 6 may also be useful.
d. Remember that positive written comments are required as well as marks. These should be recorded in the markbook too.
e. When assessing end-of-year work, be very clear about the criteria students are expected to meet (i.e., the characteristics of an adequate answer) and conscientiously try to avoid potential sources of inconsistency.
f. When part of a team of markers, work together both before and after your marking to remove idiosyncrasies due to different 'standards'.
g. Multiple marking of the same papers is a sound practice for especially important exams or assignments.

IV Some pitfalls to avoid:

a. Try not to focus solely on the 'mechanics' of writing. Excessive correction of pupils' written work is unlikely to induce better writing.
b. There is no need to assess everything that is written in the classroom. Formally evaluate only work considered by the student to be a finished effort. Allow students to revise, especially their coursework.
c. Do not mystify students by adopting marking 'standards' unknown to your pupils. Make your expectations known; make them reasonable!
d. Make sure questions and topics are not ambiguous; if they are, make allowance for this in your evaluations.

Notes

The quotation in the introduction is from the Ontario Ministry of Education, *Evaluation and the English Programme*, 1979, p.15.

Teaching and the Writing Process

Research on the writing process includes:

Cooper, C.R. and Odell, L. (eds.) *Research on Composing: Points of Departure*, N.C.T.E.: Illinois, 1978.

Emig, J. *The Composing Processes of Twelfth Graders*, N.C.T.E. Research Report, 13: Illinois, 1971.

Graves, D.H. 'An Examination of the Writing Processes of Seven Year Old Children', *Research in the Teaching of English*, 9,3, 1975, pp. 227-241.

Graves, D.H. *Balance the Basics! Let Them Write*, Ford Foundation: New York, 1978.

Collins and Gentner's study, 'A Framework for a Cognitive Theory of Writing' can be found in:

Gregg, L.W. and Steinberg, E.R. *Cognitive Processes in Writing*, Erlbaum: New Jersey, 1979.

Some useful references on assisting the developing writer are:

Hillerich, R.L. 'Developing Written Expression: How to Raise—not Raze—Writers', *Language Arts*, 56,7, October 1979, pp. 769-777.

Stibbs, A. *Assessing Children's Language*, Ward Lock Educational: London, 1979.

Thornton, G. *Teaching Writing: The Development of Written Language Skills*, Edward Arnold: London, 1980.

The importance of revision as part of the writing process is discussed in:

Calkins, L.M. 'Children's Rewriting Strategies', *Research in the Teaching of English*, 14,4, December, 1980.

Graves, D.H. 'What Children Show Us About Revision', Research Update, *Language Arts*, 56,3, March, 1979.

While Mina Shaughnessy discusses the kinds of mistakes made by first year College students in New York, many of her observations and recommendations are particularly useful for teachers of all levels in New Zealand and Australia. They are presented in:

Shaughnessy, M.P. *Errors and Expectations: A Guide for the Teacher of Basic Writing*, Oxford University Press: New York, 1977.

What Influences the Awarding of Marks?

Research on this topic is extensive, particularly on the reliability of essay markers. The references given here are a tiny selection only.

The effects of 'surface' features on markers, for example, can be found in:

Briggs, D. 'The Influence of Handwriting on Assessment', *Educational Research*, 13, 1970, pp. 50-55.

Marshall, J.C. and Powers, J.M. 'Writing Neatness, Composition Errors, and Essay Grades', *Journal of Educational Measurement*, 5, 1969, pp. 97-101.

For the effects of different marking criteria, see these early studies:

Diederich, P., French, J.W. and Carlton, S. 'Factors in Judgements of Writing Ability', *E.T.S. Research Bulletin*, 61.65: Princeton, N.J., 1961.

Remondino, C. 'A Factorial Analysis of the Evaluation of Scholastic Compositions in the Mother Tongue', *British Journal of Educational Psychology*, 29, 1959, pp. 242-251.

A useful summary of inter-marker and intra-marker reliability (i.e., marking differences in the same person), with special reference to essays is:

Cowie, Colin. 'Using the Essay as an Assessment Technique', *set 77*, no. 1, NZCER, 1977.

Assessment Techniques

(i) The analytic marking schemes described can be found in:

Diederich, P. *Measuring Growth in English*, N.C.T.E.: Illinois, 1974.

Evans, P.J., Brown, P. and Marsh, M. *Criteria for the Evaluation of Student Writing, Grades 7 & 8, A Handbook*, O.I.S.E., 1977.

(ii) For holistic marking see:

Cooper, C.R. 'Holistic Evaluation of Writing' in *Evaluating Writing: Describing, Measuring, Judging*, edited by C.R. Cooper and L. Odell, N.C.T.E.: Illinois, 1977.

Greenhalgh, C. and Townsend, D. 'Evaluating Students' Writing Holistically — An Alternative Approach', *Language Arts*, 58,7, October 1981, pp. 811-822.

(iii) A standard reference for teachers interested in essays as an examination technique is:

Coffman, W.E. 'Essay Examinations' in *Educational Measurement*, edited by R.L. Thorndike, American Council on Education: Washington, 2nd ed. 1971.

(iv) The importance of comments teachers make is discussed by:

Searle, D. and Dillon, D. 'Responding to Student Writing: What is Said or How it is Said', *Language Arts*, 57,7, October 1980, pp. 773-781.

Wade, B. 'Responses to Written Work: The Possibilities of Utilizing Pupils' Perceptions', *Educational Review*, 30,2, 1978, pp. 149-158.

Wade also cites the findings of Page and Rosen.

Item 7

Observation

The Basic Techniques

set research information for teachers
number one 1985

First issued in set, research information for teachers, no. 1, 1985

© New Zealand Council for Educational Research and Australian Council for Educational Research.

Cover illustration by Karin van Roosmalen

Printed by Lithoprint (NZ) Ltd.,

Observation:
The Basic Techniques

Bruce McMillan and Anne Meade
Otago University *NZCER*

Introduction

Most of what we know about children comes from watching carefully what they do. Proud parents entertain their friends — or bore them out of their minds — with the latest tales of their offsprings achievements. Just as frequently we discuss some 'problem', often asking a simple question such as, 'when did *your* child begin to walk?' or, 'when *do* they stop sucking their thumbs?'

Teachers often ask more complex questions such as how children of a certain age can be expected to interact with each other, or how they learn difficult concepts. Checking with other experienced people helps. Reading textbooks helps. But watching children is both more interesting and more reliable.

Observing and recording what children do sounds a simple process, and most of the time it is. But when the observations have to be used for an important purpose we find that different people see different things. This is quite usual. Think of a car accident and the evidence given in court — it seldom tallies exactly and if it did there would be suspicion of collusion amongst the witnesses! Similarly some fighting between two small children will be seen differently by the two mothers, by the preschool supervisor, and by the children.

What gives us our own peculiar and therefore somewhat unreliable view of what happens? Expectations are a common obstacle to good observation. Just because David was in a fight yesterday doesn't mean he must have started the one today, but it might make us inclined to think so. There are plenty of other obstacles. Scientific observation has to be deliberate and systematic, carried out with care and proper preparation.

The Uses of Observation

Observing children carefully and systematically enables us to go beyond guesswork or assumption, or bias. But we need a variety of techniques: each has its own range of uses. Some of the uses are suggested here.

1. We can use observation to describe the behaviour or characteristics of a particular child.

Many statements about children are simplistic generalisations, which just put a 'label' on a child: they do not describe him or her. 'Mary is terribly shy' or 'Michael is hyperactive' are examples. The labels 'shy' or 'hyperactive' tell us very little about the child. They say nothing of the good things they may do; nothing about the circumstances in which Mary may be shy, or may be quite happy to interact with others; nothing about the range of activities Michael does engage in, or how long he spends at them. When we have carried out systematic observations of a child, *then* we are entitled to draw the evidence together, and say, for example, 'Mary spends only a small part of her time at preschool playing co-operatively with others, and tends to go elsewhere if there are more than two children playing where she has been'. Or we may conclude that 'Michael stays at an activity for an average of two minutes only, and seldom talks with an adult while he is playing'. In both these cases, observation allows us to describe the children more accurately, and suggests aspects of their behaviour which could be attended to more carefully.

2. We can use observation to monitor a child's development.

Relatives or friends who see children only at irregular intervals frequently comment on how much they have grown, or changed. But those who are in constant contact with the child can fail to 'see' such developments, for they often involve slow processes of physical growth, or the acquisition of social skills, or developing thinking abilities. It is a relatively simple matter to measure a child's height every month or so. It takes rather more skill to observe and record other developmental progress. But it can be done, and can provide important information to those who are responsible for helping this progress.

3. We can use observation to examine children interacting.

Sometimes we need to know about the group of children, rather than about any particular one in the group. How does the group decide what to play? How does the group develop an idea so that the whole nature of the play,

or other activity, changes? How does the group set about including, or excluding, some particular child? When does a 'friendly tussle' become an angry fight? In all these cases we need the skill of looking carefully at all that is going on, rather than focusing only on one or two children. The results of such careful observations easily justify the time spent, for we emerge with a much clearer picture of what is happening.

4. We can use observation to examine particular play or learning.

Sometimes, we need to look at an activity or a particular situation, rather than any one child or group of children. In a preschool, for example, we may find some people complaining that one corner is never used, or another one is always left untidy. It is only too easy to fix the blame for such circumstances on to the things which immediately take the eye. A more effective procedure is to observe carefully, for some time, to check exactly what does happen, who comes to play there, and the sequence of events as they play.

5. We can use observation to check set activities, programmes, or changes in them.

It is very important that when changes are introduced there is some way of monitoring the effects they have. It is appropriate to have a set of observations before and after, or perhaps during the change. When a new item of play equipment is introduced, for example, does it take children away from other, equally valuable, activities? If so, how long does the effect last? Or, when the layout of a centre is changed, do children begin to move differently between activities? Do they spend more or less time on some of them? Only when questions such as these have been answered, do we have the information on which to base a judgement, and conclude that 'this' equipment is better than 'that', or 'this' arrangement better than 'that'.

When parents teachers and researchers are setting out to understand children and the environment in which they live careful, systematic observation *is the most important tool which they can use. It involves planning, not just casual observation. It involves careful thought about the purposes for which the observations are to be used. The more complex the purposes, the more time and effort is required in planning and carrying out the observations.*

Observation Techniques

1. Diary description

(This approach is sometimes called *anecdotal recording*)

The diary description is a fairly informal account of some aspects of one child's development. The recorder usually makes notes of any events which happen to interest him or her, over a period. Parental pride may encourage us to record our child's first language, or motor skills such as walking, for example. Or a teacher may make a few quick notes about a child's first day at preschool, or school, and occasional days following that. It probably means that these occasional diary entries are made whenever the right mood happens to strike the observer, rather than on any systematic basis. Because of this, the diary record can be quite inadequate as a sound description of the child, or any specific aspects of his or her development. The observer can be biased towards a certain kind of situation, or perhaps select only some particular things to record. This means that usually no great use can be made of the observation records.

Nevertheless, diary descriptions do have their uses. The observer is very interested in some aspects of the child's behaviour, and prepared to take the time to note down impressions. Even beginning such a simple recording can help to sharpen the observer's awareness of what is going on. That can, in turn, lead to the realisation that different types of interesting behaviour are appearing, or that there may be relationships between some of the things that have happened.

For example, the parents' diaries interspersed through Margery Renwick's *To School At Five*. (Wellington, NZCER, 1984), illustrate how such records can capture change. The children were experiencing a major life transition and the diary descriptions suggest the contexts that go with relatively smooth transitions.

Thus, diary descriptions can be the springboard to further, more systematic observations. These observations will be specifically aimed at finding out the answers to the questions that arose from whatever we happened to notice.

Diary descriptions could obviously be helped by photographs: the family photo album can be a record of the development of children in a family! Film can have the same use. But written records are most common, and have the advantage that only pen, paper and a little thought are necessary. Charles Darwin observed and recorded the behaviour of his son in this way because his scientific background helped him to see the value of careful description.

Here are some more examples. The first two are entries from a diary kept by a mother who was particularly interested in language, and who simply wanted to keep a record of her child's progress in this area.

> **B**'s day was at an end. Bathed and fed he sat wrapped in a shawl on my lap to have his 'evening talk' with me. All at once he looked intently at the wall at my back. The evening sun lay on it in a broad golden band mirroring the window and latticed by the black shadows of moving leaves. Intently he watched it all, then looked up at me with a smile; he uttered a delicate sound, and looked back. With that sound he spoke to me of all the loveliness he was seeing, and wanted to know whether I saw it too.
>
> February 1st: A good while ago, **B** used to bring me his shoes when I pointed to them and said 'bring me the shoes!' This and similar occurrences might have suggested that he understood my utterance. But it is quite possible that my pointing to the shoes and saying something or other was sufficient to suggest it to him. But now, **B** definitely understands whole phrases. When I say to him 'we will go to the bathroom', up he gets, and goes to the door in order to patter along the hall-way and play with the empty shampoo bottles in the bathroom.

Another example preserves the amusing and perhaps half-understood ideas of a small boy. It was first published in 1892.

> **R** came into the house eating a horse-chestnut.
>
> **Grandma:** Well, **R**, if you eat that horse-chestnut you'll die and go to heaven with your mother, and then I shan't have any **R.** (His mother was dead.)
>
> **R:** Well, I'll go out and get four horse-chestnuts, one for grandpa, one for you, one for Aunt H and one for me. I'll eat mine first, then I'll die and go to heaven first. Then grandpa'll eat his and he'll die, then you'll eat yours and we'll all be up there together. Won't that be nice, grandma?

(Cited by Herbert F. Wright, 'Observational Child Study', in Paul H. Mussen (Ed.), *Handbook of Research Methods in Child Development*. New York: Wiley, 1960, p.83.)

2. *Running record*

A running record provides a description of one child's behaviour over a period. It is one way of building up a careful description of what the child does, but depends on the record containing a good description of the environment as well. In other words, it attempts to provide an account

of what the child does from moment to moment, in a particular setting.

Usually, we attempt to record as much of the behaviour as possible. That is difficult, and we find that about 15 to 20 minutes is as much as we can do at one period. However, a really useful record can be built up by doing a number of observations, with either a few minutes, or hours, or days in between.

The major advantage this technique has is that, by trying to note down on paper (or into a tape-recorder) *everything* that happens, we can see the complex network of interactions a child has with others, and with the environment. The major disadvantage is that so much is likely to happen that we become selective, or begin to lack precise descriptions of what goes on. In such cases, running records can be misleading, and more precise techniques must be brought into use.

Here is a small example of one four-year-old girl's behaviour, over a four-minute period. It tries to convey the general picture, but some of the detail of the picture 'washes out', since there is so much to try and record.

> **A** at paste table with mother — mother leaves table and **A** gazes around dreamily at the other children. **A** now leaves the paste table without doing any work, she goes over to dough table, but there are no empty seats and she doesn't know where to go. She is rather uneasy — looking around her. Mother then takes her by the hand into the dolls' corner where she stands beside the teaset table just watching the other children play. While still standing at table she pulls out the chairs and pours the (empty) teapot into the cups — leaves and wanders around the dolls' beds, picks up some dolls' clothes from the floor then just throws them down again, gives the rocking bed a push as she passes then takes doll from pushchair, puts it carefully to sleep in the pram, tucking the rug gently around the doll — leaves pram and settles another doll into a bed, tidying blankets and folding sheets, then once again tidies the dolls' bed very neatly. (Solitary play.) **A** leaves corner, walks slowly towards dough table all the time sucking her finger (still no room) still sucking finger she moves slowly over to paste table — takes some black paper over to table and starts to paste.

By opening up a wide range of *possible* behaviour (or events) to observe, running records can be most valuable.

Pamela Kennedy questioned whether entry to school at five is a transition or trauma (*Early Childhood in New Zealand*. Second Early Childhood Care and Development Convention, 1979). Through running record observations of children around five years old in pre-school and junior class settings she was able to show how children in the two settings

had different opportunities for motivation, concrete experiences, social interaction and getting things in balance. These are aspects of growth which Piaget considered important for children if they are to develop their concrete operational thought and reasoning. Thus, running records of several children provided for Pamela Kennedy generalized data on child development at a specific age.

3. Time sample
(of individuals or the use of an activity)

This is a development of the running record. A 'beeper' or some timing device gives the observer a 'beep' and only what is happening at that moment is written down. Then, after a gap, another 'beep' and what is happening is noted again. Thus the behaviour is 'sampled' at pre-arranged times, usually at 1, 3, 5 or 10 minute intervals.

In this way, time-sampling is rather like taking a 'slice' of time out of a running record. Or, it is like taking a single frame out of a movie film at regular intervals, so that each frame can be looked at more carefully. When sufficient samples of time have been taken, we can start to draw the threads together.

With time sample observations of an individual child we may, for example, discover that whereas we thought Mark played with other children quite regularly, it just so happened that we noticed him only when he was playing with others. Now that we have looked more systematically at his play, we realise that he was playing with other children for only two of the twelve samples we took over an hour: that is hardly regular social contact.

With time sample observations of an activity area, we may, for example, discover that blocks are seldom used until after morning tea. We realise that it is not until the children have sat nearby that they remember the block corner tucked behind shelving. By use of the same technique, we may discover that as the number of boys using the blocks rises (through any hour of regular observations), the number of girls using them declines.

A time sample of all activity areas during free play can provide information on how children are dispersed, how popular different activities are, and the patterns of play at the beginning or end of the sessions. As the number of cases or activities you want to observe increases, so the information to be gathered needs to be made (or kept) simple.

One value this technique has is simplicity. A wrist-watch with a sweep-second hand or running seconds is all that is actually needed for timing. By 'sampling' times during the preschool session, or during the day at home, or even over periods of days, an accurate record can be built up.

4. Time sampling using categories

This is similar to number 3, except for two considerable refinements. Recording is reduced to writing a simple code or putting a check mark alongside a list at the sampling moment. It registers just the presence (or absence) of a specific behaviour. With such a simple task for the observer it becomes possible to observe groups of children as well as individuals. This procedure is widely used in behaviour analysis studies where, for example, it can provide systematic records of the frequency of 'problem' behaviour. This can be done during baseline surveys and during treatment.

It is at first glance, quite a simple procedure, for it avoids some of the problems you get when long descriptions of behaviour have to be written down. But this simplicity is deceptive. *The category* used for the behaviour you want to observe *must be very carefully defined*. A category label such as 'creative play' for example, means virtually nothing, until you have gone through the difficult task of defining it, and giving examples of the kinds of behaviour you could or could not code in this way.

An example of a simple recording chart, in which appropriate (A) or inappropriate (I) behaviour by three children was recorded is shown below. Note that it is very easy to see what the average amount of 'appropriate' behaviour is. For Harry and Tom it is 50% of the times, but for Dick, it is 70%. What 'appropriate behaviour' is, must, of course, be very carefully spelled out.

Time	Harry	Tom	Dick
9.20	A	I	A
10.05	A	A	I
10.15	I	I	A
11.00	A	A	A
11.20	I	I	A
11.45	I	A	I
1.05	I	I	A
1.30	I	A	I
2.20	A	A	A
2.50	A	I	A

(From E.L. Glynn, 'Introducing Behaviour Analysis,' in New Zealand Educational Institute Yearbook No. 4, *Children's Behaviour — its modification by teacher, parent and peer*. Wellington: N.Z.E.I., 1975, p.35.)

Where a trained observer can be present in a class or centre, it is possible to have much more careful attention paid to the numbers (and percentages)

of children who are engaged in the various classroom activities. In this way, an accurate picture of the pattern of interest and attention throughout the day can be built up, and steps taken to remedy any shortcomings in the programme. (For further reading on this, see the paper by Glynn referred to above; or Carol A. Cartwright and G. Phillip Cartwright, *Developing Observation Skills*. New York: McGraw Hill, 1974; or Todd R. Risley and Michael Cataldo, *Planned Activity Check*. Kansas City: Center for Applied Behaviour Analysis, 1973.)

5. Interval recording

This is a further refinement of time sampling, with the teacher or researcher focusing on (i) one person only, (ii) a number of categories of behaviour. In this case, a simple 'yes' or 'no' is recorded to the question, did the behaviour occur? For example, it is common to observe for 10 seconds, looking for any behaviour fitting one or more of the categories. Then there is a 5-second interval for making the check marks, followed immediately by the next 10-second observation period. And so on for 15 minutes or so. An example: Marsha Weinraub and Jay Frankel ('Sex differences in parent-infant interaction during free play, departure, and separation', *Child Development* 1977, v.48, pp.1240-1249) were able to show that:

> Parents talked to, got down on the floor to play with, and tended to share play more with same-sexed than opposite-sexed infants (aged 15 months to 21 months) . . . When infants were close to their mothers, mothers were more likely to look, vocalise, touch, sit on the floor and share play with their children. This was not true of fathers.

How did they know this?

Parent-infant behaviours were observed from behind a one-way mirror . . . Using checklists, the occurrence or non-occurrence of particular behaviours within 5-second intervals was observed; a 3-second interval was used to record the data. An audio recording delivered to the observers' headphones cued the observer when to observe and when to record. Parent behaviours included looking at the infant, reading magazines, and sitting on the floor.

A New Zealand example is Anne Meade's *Team Teaching in New Zealand Early Childhood Centres* (Wellington, NZCER, forthcoming). The observers were able to show that 'there was less adult-child talk than most early educators expected, [and] adults getting down to child level was associated with more sustained conversations . . . . Trained staff did more to foster children's learning through talking to them and through play involvement than parent helpers', (p.4).

In this case, the observers shadowed on-duty adults and when a beeper gave the cue (every 30 seconds) the observers checked which of 17 types of behaviour occurred in the 3 seconds following. The categories covered adult-child talk, affective (emotional) behaviour, play involvement and adult-adult interactions.

It can be done, but it is difficult, to use this technique without some specialised equipment to produce the sound which signals when to switch from observing to recording. Electronic timers are inexpensive but need modification to get them to beep repeatedly. However, this is a technique you should think of carefully: it combines (i) the sampling of behaviour needed for systematic recording with (ii) the narrow focus on a few specified types of behaviour required for detailed analyses. It could be adapted to use more simply: a timekeeper whispering the 'beeps' in the observer's ear could do the job. The schedule for interval recording could look like this example, (a tick (✓) indicates that the behaviour *was seen* in that interval):

Time Interval	Category of Behaviour 1	2	3	4
1	✓			
2		✓	✓	
3	✓			
4	✓	✓	✓	
5	✓	✓		✓
etc.				

6. *Event recording*

The focus of attention is an event, for example, the occurrence of a certain sort of behaviour. Time intervals between events are *not* important. This approach can be as simple as recording the number of times a child uses a certain word, or plays co-operatively with others. Events can be recorded using pencil and paper, a wrist 'golf-counter', a knitting counter, by electronic event recorders, or even by transferring pebbles from one pocket to another.

Here is an example of a complex procedure: one study showed how much preschool children reinforce each other's social interactions. It is no surprise, of course, to find that someone being nice or nasty to you tends to make you nice or nasty back to them: the important point is that the number, kind and frequency of specific events were revealed. (See Michael P. Leiter, 'A study of reciprocity in preschool play groups', *Child*

Development, 1977, v.48, pp.1288-1295.)

In his study, Michael Leiter observed for 15-minute periods, for a total of 40 hours; with several carefully-defined categories:

. . . all social initiations made by a target child are recorded, along with the immediate social response to the initiations . . . . The coding record was made on blank sheets of paper on a clipboard with a felt pen . . . . Child-child interactions were recorded for 16 subjects during free play time at their preschools . . . . Each child was observed for 20 15-minute intervals, yielding a 5-hour record of behaviour for each child. It was found feasible to observe two children simultaneously.

Event recordings do not have to be as complex as this. A simple technique would be to watch children's play at a particular activity, and note, say, the frequency of conversations or physical contact. A New Zealand example is Elizabeth Connelly's study of the sandpit and block areas in a number of Auckland kindergartens. She noted rejection events and found that only a small proportion of girls were actually rejected from these areas during the fourteen sessions observed. This data led her to look for other explanations for the girls' low participation rate. (Massey Certificate in Early Childhood Education Special Topic, 1974.)

7. Duration recording

The aim of this procedure is to establish how long a particular behaviour lasts. A stop-watch is all that is required if a single child is to be observed, but more useful information about matters such as time spent in various activities is gained by using a prepared sheet, and fitting it in, as in this example.

Activity							**Blocks**
Child Time	9.00	9.05	9.10	9.15	9.20	9.25	9.30	Total
Mary	▬▬							5
Joe		▬▬	▬▬	▬▬	▬▬			15
Kathy			▬			▬▬		7
Mark		▬▬	▬▬	▬▬	▬▬			15

(For the results of such an exercise, see Bruce McMillan, 'One hour in kindergarten', *Preschool Education*, 1971, v.1, pp.18-20.)

A more complex variation of this schedule was used in 1983 for a study of girls' participation in play with blocks and car cases in six kindergartens

(an NZCER/Society for Research on Women in New Zealand project). Teachers' names were added as well. Thus, it was possible to check whether more girls played for longer durations when the teacher(s) were present (and, similarly, whether the rates changed when boys joined in). This study will be reported in the second issue of *set* for 1985.

A different approach to this is to study one child, with one specific category of behaviour in mind perhaps after such a comment as 'Lee is *always* fighting other children'. The example below was prompted by someone saying, 'Mary is *always* wandering aimlessly around'.

Anne Smith provides a sample schedule for Mary. Duration recording showed clearly that Mary 'wandered aimlessly' on four occasions and only one-third of her time was spent this way when calculated thus:

$$\text{Percent wandering} = \frac{\text{time wandering whilst observed}}{\text{total time observed}} \times 100$$

The observation schedule (the form filled in by the observer) included a precise definition of all behaviour which could be classed as 'wandering aimlessly'. On the schedule the observer noted the exact times throughout the day that Mary started and stopped wandering. (See Anne B. Smith *Understanding Children's Development*. Sydney: George Allen Unwin, 1982, p.45).

8. Trait rating

This approach is much less precise than many of the others, but still has a place, provided of course that it does depend on careful observation rather than 'hunch'. A child is observed, probably for a specified time, and then given a rating on a given trait, for example, is rated 3 on a 5-point scale for 'friendliness'.

Trait rating was used, for example, in a study by Walter Emmerich ('Evaluating alternative models of development', *Child Development* 1977, v.48, pp.1401-1410).

It should be noted that each rating was based on approximately 30 minutes continuous observation of a target child within a free play or small group context. Following each such observation, the observer immediately completed a rating schedule. This included 149 Unipolar Scales, explicitly defined by a manual. (p.1405)

Nevertheless, it is clear that trait rating is still likely to be unreliable (see David Y. Schulter and I. Rogis McNamara, 'Expectancy Factors in behavioral observation', *Behaviour Therapy*, 1976, v.7, pp.519-527). It

is a technique to be used very carefully and preferably only to supplement other approaches. We have very clear research evidence that even teachers, busy observing pupils every day, can, for a host of reasons, rate children's traits very inaccurately. Ask early educators who team-teach in the same centre to rate children and you will cause much debate because each will perceive the children's characteristics differently.

Some Further Considerations

1. The definition of categories

The most fundamental step in observation is specifying as precisely as possible the behaviour to be observed. Vague or generalised descriptions lead only to frustration: nobody else can be really clear about what is being observed, comparisons or progress checks are not reliable, and the observers themselves are not sure whether the behaviour fits in one category or another. 'Playing', as a description, for example, is most inadequate: 'playing with wooden blocks' is a little more precise; 'stacking wooden block on top of two similar ones' is even better. The precision required may vary, but in *any* observation there are two fundamental rules:

(a) you must start out to observe and record only behaviour which *can* be clearly seen and/or heard;

(b) you must differentiate between similar types of behaviour if they are likely to be confused. Consider, for example, physical contact between young children which may be quite different depending on whether pushing or hitting is involved.

2. The number of categories

The number of categories you can handle reliably varies according to the clarity of definition, the technique for recording and your experience. Observations are more likely to be accurate and reliable if there are few categories.

How many is too many? Studies of behaviour analysis use about four to eight categories; for example, a classroom study used one 'on-task' and six 'off-task' types of behaviour. (J.D. Thomas, F. Pohl, I. Presland, and E.L. Glynn, 'A behaviour analysis approach to guidance', *New Zealand Journal of Educational Studies*, 1977, v.12, pp.17-28.) A number of child development studies report observations with many more. For example

the very important study of mother-child interaction by Allison Clarke-Stewart, observed 26 maternal, and 23 infant types of behaviour, using an event recording approach. (K. Allison Clarke-Stewart, 'Interactions between mothers and their young children: characteristics and consequences' *Monographs of the Society for Research in Child Development*, 1973, v.38, nos.6-7.) With so many categories to be observed, there needs to be a great deal of training and practice for observers to ensure satisfactory levels of agreement between different observers.

3. Obtaining accurate and reliable observations

The most important consideration when undertaking observation studies is making sure that what is being observed is important, and likely to be of value to the teacher, parent or whoever is undertaking the study. In some cases, then, the question of accuracy does not arise: the diary description given earlier is one parent's record of some aspects of her child's development, and its value lies in being a personal record. It makes no attempt to be a thoroughly complete documentation of the whole of **B**'s development. But where observational records are to be used for specific purposes, every attempt must be made to reach a level of accuracy and reliability.

Accurate observations require clear, unambiguous definitions for it must be possible for the observer to record exactly what the behaviour is. If 'aggression' is to be recorded, for example, it may be that accidental contact, or jostling, needs to be separately considered. Observers must 'get it right' each time.

Reliable observations require that whoever is observing records the behaviour consistently across several sessions. You can be reliable, but wrong, of course. However, that is easier to fix up than being unreliable.

The issue of accuracy and reliablity is very large, and cannot be considered in full here. Teachers or parents usually want to use observation as a basis for changing something. It is sufficient for them (i) to practise until they are confident they know the techniques, (ii) to check their definitions of each type of behaviour with others who are interested, (iii) to check the extent to which they and another competent observer agree about how each type of behaviour is to be recorded before beginning to record, and (iv) to have other occasional checks during the course of any large series of observations. Inter-observer agreement, in its simplest form, means that two observers independently make a record of the behaviour, and then check the extent to which they agree. Where categories of

behaviour have been checked, the task is easier, as the agreements can be counted. A common procedure is to apply this formula:

$$\frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}} \times \frac{100}{1}$$

This gives a percentage agreement. Usually 85% or so is considered an acceptable level of agreement. Where observations are part of a research programme much more needs to be considered, but for classroom or pre-school practices, the above should be sufficient.

4. The effect of an observer's presence

It is generally agreed that being watched may make for unusual behaviour. This is more likely to be the case when, for example, in a family home the observer is unknown to either the child or the parent. It is much less likely if the observer is in a class or centre that adults often visit. However there is no way of telling how much difference observers make in different settings, and so only general guidelines can be given on the topic.

First, it is important that the observer remains as unobtrusive as possible. This may well conflict with the necessity of being close enough to see and hear the behaviour which is being observed, but 'unobtrusiveness' is an attitude as much as anything. It is clear that an observer cannot both interact with the child (or person being observed) and then change back to the impartial role. Observers should also avoid the mistake of paying obvious attention to any particular behaviour: it is likely that a sudden show of activity on the part of the observer when aggression (or whatever behaviour is being observed) occurs, will increase the chances of that behaviour occurring again.

Second, it is preferable for observations to be spaced over a reasonable period of time. It is probable that any effect arising from an observer's presence will be greatly decreased over time, as the subjects become accustomed to his or her presence.

Third, it is also probable that the effects are much reduced when the subjects are younger: babies and infants are unlikely to change their behaviour solely in response to the presence of an observer. However, they will respond to many 'irrelevant' changes. If a young child's behaviour is being observed in the presence of his parents, for example, if the parent changes his or her patterns of interaction, that will almost certainly change the picture for the child's behaviour as well.

Finally the whole context in which the observations are being conducted

75

needs to be considered. Children in primary schools in the main cities are notoriously familiar with the visits of student teachers, and tend to ignore them. That may well be an asset for the observations the students have to do. On the other hand, children in playcentres and playgroups are used to having several parents around during sessions, and know that they can call on their help, or simply converse with them. Such a familiarity makes it more difficult for a playcentre supervisor to be unobtrusive whilst observing.

Another important consideration is the use to which the observational material is to be put. In the playcentre context again, this is easily accepted as 'Mrs Smith is doing her observations — again'. Little further explanation is required, either by the children or by the other parents. However, where a different kind of observational approach is used, or where any observations are a novelty (as in research observations in a family), it is very important that the reason for the observation is explained, and reassurance given that no particular change in behaviour is required.

5. Some ecological considerations

All observational reports need an account of the various factors which may affect the child or the setting. Thus the date and time of the observation, and a brief description of the setting (room arrangement or play area) should be attached. The weather may also be an important factor, especially if specific kinds of play are being observed over a week or so.

Where the observations are being used to check the effects of changes in teaching or setting, these ecological considerations are terribly important, for otherwise there is the danger that changes may be wrongly attributed to what the observer wants to see. Even the number of adults present at a certain time may radically alter patterns of play or other activity and so these factors should also be recorded.

A number of writers insist that observations can only be undertaken in the 'natural environment' and that no specific stimulus should be introduced. But as we have noted, the presence of an observer has already changed the environment. So does the time of day, in some cases. (For example, observations of very young children are likely to be greatly different if they take place in the morning, and the late afternoon.) Many observational reports are interested in what happens when some specific play material is introduced, or when television programmes are shown, or when there is one of any number of contrived stimulus situations. Provided these stimuli are referred to in reporting the observation, there

is no reason why they should not be used.

6. *A note about ethical concerns*

Observations should only be undertaken when two conditions are met:

1. The permission of the person observed (or parent if necessary) has been obtained.

2. The material recorded as a result of the observation is kept confidential.

In both cases, this means that the child, parent or teacher, needs to have the reason for the observation spelled out in a way that they can readily understand; and proper safeguards for confidentiality should be explained before the observation is done.

If you fail to get permission and do not keep the results confidential it is not surprising if the people you want to observe refuse to let you, or others, ever observe again.

Video Recording for Observing

If you have a well made video recording many of the techniques described above can be applied to analyse the behaviour you have captured. And you can use first one technique and then another. This takes away some of the criticisms about how selective observers' views can be.

The use of a video recorder, as a technology to aid observations, is a major topic in itself and a full item in *set* is planned for a future issue.

The Authors

Dr Bruce McMillan is a Senior Lecturer in Education at Otago University, Box 56, Dunedin, New Zealand.

Dr Anne Meade is a Research Officer with the New Zealand Council for Educational Research, Box 3237, Wellington, New Zealand.

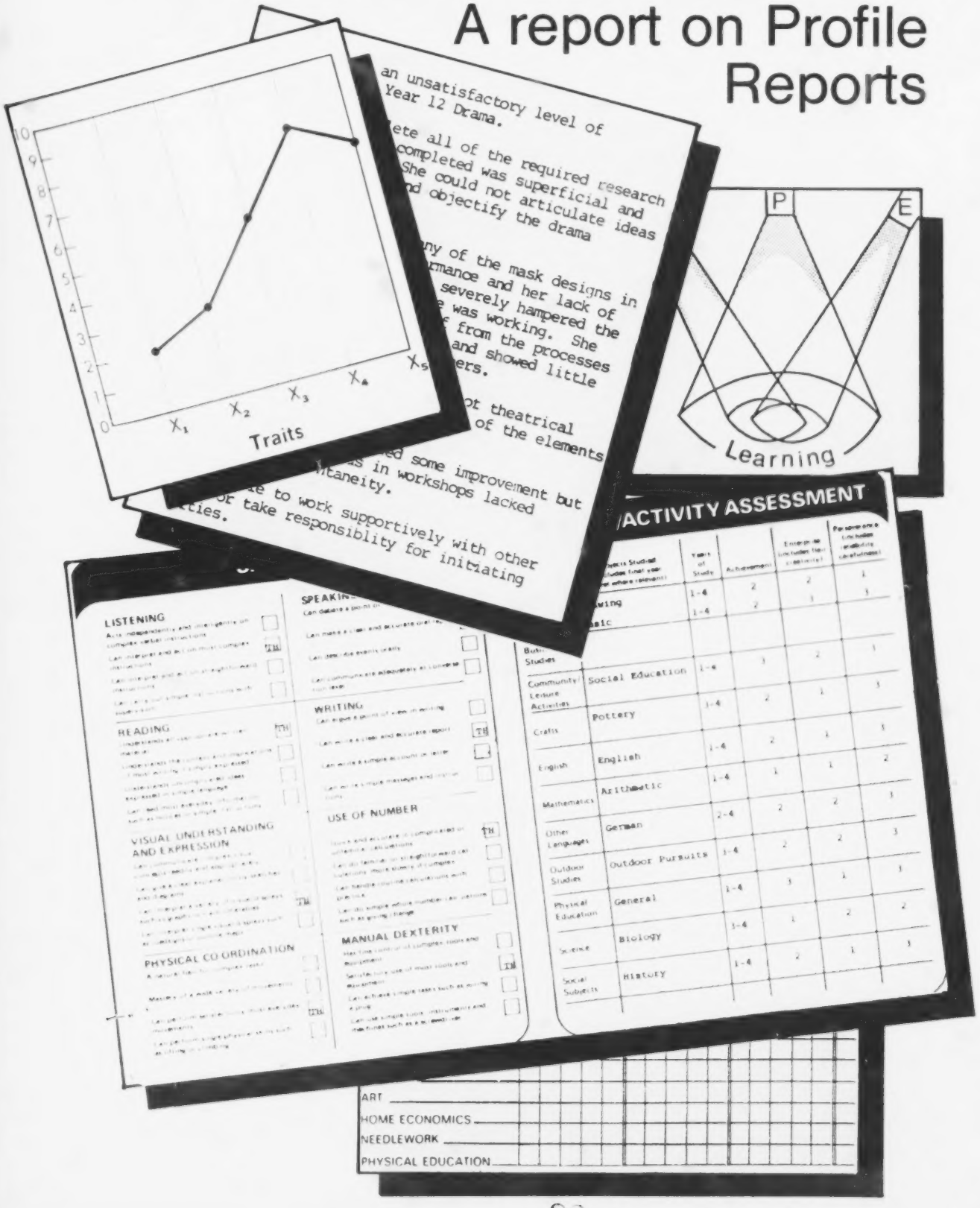
An NZCER Information Service

81

set research information for teachers
number one 1987

ONE EXTREME TO THE OTHER:

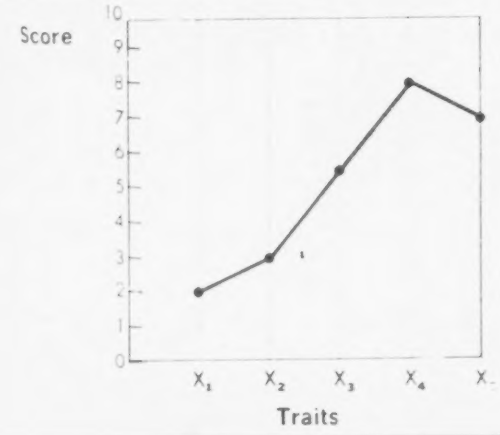
A report on Profile Reports



One Extreme to the Other: A Report on Profile Reports

By Graeme Withers
ACER

James k's Profile on 5 Traits



Sue k's Profile in Drama

Sue has shown an unsatisfactory level of achievement in Year 12 Drama.

She did not complete all of the required research and work that was completed was superficial and limited in scope. She could not articulate ideas clearly or analyse and objectify the drama experience.

She failed to complete any of the mask designs in connection with the performance and her lack of commitment to this project severely hampered the other students with whom she was working. She consistently isolated herself from the processes involved with the performance and showed little flexibility in working with others.

She attended the required number of theatrical performances but her understanding of the elements of the art was extremely limited.

Her expressive skills showed some improvement but her exploration of ideas in workshops lacked imagination and spontaneity.

She was unable to work supportively with other students or take responsiblity for initiating activities.

Both these profiles follow someone's definition of what a 'profile' is. The most common 'profile', however, describes the PERFORMANCE of ONE student, in SEVERAL subjects, during ONE year.

The enormous differences between what one teacher means by a profile and what another means leads to confusion amongst teachers, parents, employers, and officials. This **set** item will display some of the different types of assessment all flying the banner PROFILE, and make recommendations about them, and their use.

What the Profilers Practice

Just as different schools and colleges produce extremely different pieces of paper at the end of their profiling so do the philosophies of the profilers differ. Two sets of ideas are at work.

1. **What ATTRIBUTES should a profile have?**
   Some will say a true profile is no more than a set of co-ordinates tightly described.
   Others will say that a true profile is a descriptive statement in words, and unquantifiable.
   Yet others will say a true profile lies somewhere between these extremes.

2. **What FUNCTIONS should a profile have?**
   Some people will stress competitive practices
   - administration (school decision making, record keeping)
   - selection (employers, further education)
   - guidance (diagnostics: counselling)
   - information (to students, parents)
   - motivation
   - discipline

   Others will stress non-competitive practices
   - enabling (better curriculum and course planning)
   - developmental (a learner's knowledge and sense of autonomy)
   - co-operative (interpersonal learning and relationships)
   - assistance (to the learner about his/her learning)
   - placement (courses, jobs, further education)

In the descriptions of seven basic types of profile which follow you will see these differing philosophies about the attributes and functions of profiles coming through. We will leave judgements about them till later.

Seven Profile Types

1. Feed-back assessment and reporting about courses

In some educational systems, a set of trial examinations is conducted at the end of the first or second terms to give a percentage score for each subject. These scores are not counted towards eventual success or failure but are expected to help motivation, guidance and discipline (see the competitive functions listed above.) At a basic level, the results provide feedback for continuous improvement and revision.

Figure 1 comes close to being a good example. It is taken from the third form report certificate of a New Zealand school. This frame is accompanied by similar frames for each of the other core subjects, and sections (allowing comment only) on each 'elective' activity. The whole document is a profile procedure; so too is Sue K's *Profile in Drama*, on page 1, a profile of English skills. The value of the report from a formative feed-back point of view will depend on the nature and content of the comments: how student-oriented, positive and specific they are.

An important distinction between Type 1 and other procedures is made by the school in introducing the report:

> This report is intended to advise parents and to assist pupils in their academic, personal and social growth. In addition to the subjects listed on the inside of this report, all pupils take seven short courses.
>
> This report is not designed to assist employers in selection procedures. Pupils seeking employment should apply for a College Record and Reference Certificate which summarises their major skills, attitudes and personal qualities at the time of writing. This should be adequate for the needs of employers.

FORM 3

First / Second HALF YEAR

	SUBJECT: ENGLISH	Grade
	TEACHER:	
	ATTITUDE:	
SKILLS	Writing	
	Reading	
	Speaking	
	Listening	
	Presentation	
COMMENTS		

Grades
- A Excellent
- B Very Good
- C Satisfactory
- D Causing Concern
- E Unsatisfactory

Figure 1: A simple subject profile

A second example for Type 1 is Figure 2, from an Australian high school. The assessment basis is the specific work completed (note the report on each of six, numbered contracts), and it contains a section outlining possible contributions by parents. If the contracts are a profile of learning, so too is this document, reporting the work completed.

2. Reporting of performance scores in course work

Figure 3, from an Australian technical high school, offers a specimen profile of Type 2. Space is left in the slanted section of

NAME: Marta

SUBJECT: English YEAR LEVEL: 8 TERM: II

		Yes	No
1	COMPLETE PROJECT/PERSONAL PROJECT Project was 7 pages, well set out, clear headings and interesting photographs. The story about grandfather explained very well what his work used to be like and how your work, when you leave school, will be very different.	✓	
2	READ NOVEL 'SUPERFUDGE' AND COMPLETE ASSIGNMENT All the work was completed but next time consider taking more time thinking about the answers. Your written answers shouldn't tell the story again but be about what you believe the story is telling us.	✓	
3	COMPLETE REGULAR SPELLING, COMPREHENSION AND VOCABULARY WORK Completed and corrected.	✓	
4	WRITE 3 STORIES AND WORK ON IMPROVING THEM All stories were improved by re-writing and you should be very pleased. Your story 'Survival' was exciting to read. You could use it for reading to the Grade 5 pupils next term.	✓	
5	KEEP A JOURNAL AND WRITE IN IT EVERY DAY You seem to enjoy writing in your journal. You might like to copy out 2 poems for the wall display.	✓	
6	CO-OPERATE WITH TEACHER AND OTHER STUDENTS, PARTICIPATE IN CLASS ACTIVITIES AND BRING NECESSARY MATERIALS TO CLASS This term has been important for you in that you often spoke in class discussions and helped organise our excursion. You must remember to complete all work on time. Next term use your diary entries as reminders.	✓	

SUMMARY: For Parents

- All assessment goals for Term II English have been completed.
- Marta improved her writing this term because she carefully followed instructions, and worked quietly in class.
- She has been a little late with some work and next term should use her diary. If the tasks are hard she should tell me.
- I look forward to discussing Marta's work with you at the parent-teacher meeting.

Figure 2: A work-completed profile statement

the table for inserting the particular sections of the course seen as relevant. There is also room (not represented in the Figure) for extended comment of a descriptive kind to summarize achievements in other sections, and on performance in the course as a whole. Space is also allocated for detail of what work was done in the course and how the judgments of quality were reached.

This is perhaps the most common type of profile. Ticking a box and entering brief commentaries are easy and quick and popular. However, how adequate is a multi-point continuum to express the full range of behaviour or attributes under each heading? That

KINGS TECHNICAL SCHOOL	REPORT TO PARENTS	DATE
SUBJECT: CLOTHING CONSTRUCTION II		FORM

This course is to extend the study of Needlecraft to a more advanced level. Students are instructed in the styling and making of fashion garments, a knowledge of the fundamentals of good design, adapting commercial patterns, fitting techniques, suitability of fabrics for specific garments and an awareness of current fashion.

COMMENT IF NECESSARY

ATTITUDE TO WORK
1. CONTINUOUS DESIRE TO WORK
2. GENERALLY POSITIVE AND EFFECTIVE
3. NOTICEABLY UNWORKABLE AND/OR DIFFICULT TO CONCENTRATE
4. UNWILLING UNLESS SUPERVISED

MANIPULATIVE or PRACTICAL SKILLS
5. HAS ACHIEVED THE NECESSARY SKILLS
6. HAS ACHIEVED MOST OF THE SKILLS
7. HAS ACHIEVED A FEW OF THE SKILLS

ACHIEVEMENT
8. EXCELLENT. SUCCESSFULLY PERFORMS ALL TASKS
9. HAS PROGRESSED VERY WELL
10. USUALLY SATISFACTORY. HAS IMPROVED IN SOME AREAS
11. OCCASIONALLY SATISFACTORY. HAS MADE LIMITED PROGRESS
12. APPEARS TO HAVE MADE LITTLE OR NO PROGRESS

Figure 3: A grid-and-comment course profile

remains to be demonstrated. Nevertheless, it does represent an improvement over a statement such as 'Clothing Construction II: pass'.

3. Whole-course reporting

Figure 4, from another Australian technical secondary school stands as an illustration of a large number of specimens which attempt to report a whole year's work by a student in a single multi-faceted statement. The lines on the right, extended, are for comment by subject and home-room teachers to amplify the judgments indicated by ticks in the boxes.

Figure 5 narrows the time focus to four weeks, and expands simple subject-based assessments to broader, inter-subject assessments. This profile derives from a highly specific Youth Opportunity Programme devised by the City and Guilds of London Institute, and has been used as a model in other places. Figure 5 is completed by shading in the relevant parts of each numbered, horizontal 'line'.

Queens Technical School

NAME _____ YEAR _____

	ACHIEVEMENT					ATTITUDE					CONDUCT				
	Excellent	Very Good	Satisfactory	Weak	Very Weak	Excellent	Very Good	Satisfactory	Weak	Very Weak	Excellent	Very Good	Satisfactory	Weak	Very Weak
HUMANITIES															
MATHEMATICS															
SCIENCE															
ART															
HOME ECONOMICS															
NEEDLEWORK															
PHYSICAL EDUCATION															

Figure 4: A grid-and-comment profile of a year's work

PILOT SCHEME

PROFILE REPORT — FOUR WEEKLY REVIEW

This profile shows the levels which have been reached during the last four weeks and the learning activities which have taken place.

ATTAINMENTS IN BASIC ABILITIES

		4 (Basic Level)	3	2	1 (High Level)
SOCIAL ABILITIES	WORKING WITH COLLEAGUES 1	Can cooperate with others when led	Can work with other members of group to achieve common aims	Understands own position and results of own actions within a group	Is an active, decisive member of group. Helps and encourages others
	WORKING WITH THOSE IN AUTHORITY 2	Can follow verbal instructions for simple tasks and can perform them under supervision	Can follow a series of verbal instructions and carry them out independently	Can carry out a series of tasks effectively, given minimum instructions	Inspires confidence in those in authority and communicates well with them
	SELF AWARENESS 3	Is aware of own personality and situation	Can determine own strengths, weaknesses and preferences with some guidance	Has good basic understanding of own situation, personality and motivation	Has a thorough understanding of own personality and abilities and their implications
'ON	TALKING AND LISTENING	Can hold conversations with workmates, face-to-face or by 'phone. Can take messages	Can follow and give simple descriptions and explanations	Can communicate effectively with a range of people in a variety of situations	Can present a logical and effective argument. Can analyse other's arguments

An similarly for communication, practical and numerical abilities, and ....

DECISION-MAKING ABIL	INFORMATION SEEKING 12	...ormation with ...ice from supervisor	...d sources ...un	...nation fr... ...ources	Show... ...eeking and gathering ...nformation from a wide variety of sources
	COPING WITH PROBLEMS 13	With guidance, can cope with simple, everyday problems	Can cope with complex but routine problems. Seeks help if needed	Can cope with unusual problems by adapting familiar routines independently	Can offer sensitive and effective help to other people facing problems
	EVALUATING RESULTS 14	Can assess own results with guidance. Asks for advice	Can assess own output for routine tasks independently	Can assess own performance and identify possible improvements	Can identify other's difficulties and so help to improve group performance

N/O — No opportunity to assess.

Name of Trainee D_____ T_____

Name of Scheme *YOP Community Playgroup*

Period covered by this review *1 Month* *26.2.82 — 23.3.82*

Figure 5: Profile of performance on a whole course

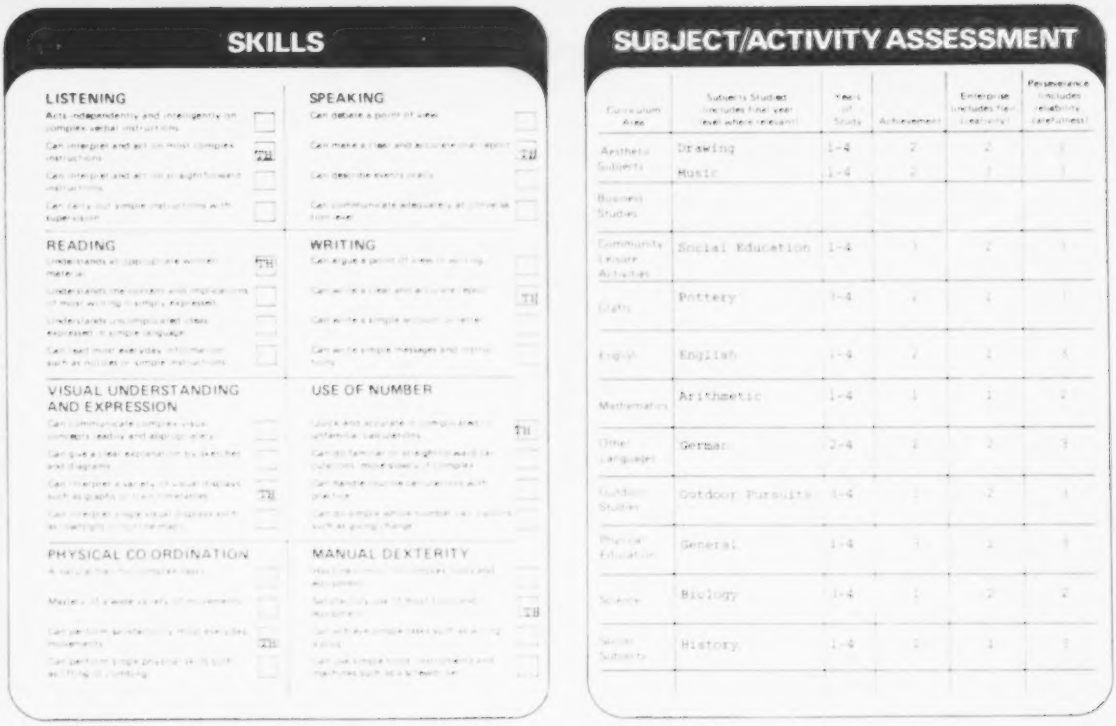


Figure 6: Another profile of performance on a whole course

A fourth specimen (Figure 6) comes from trials in 1977 by the Scottish Council for Research in Education. Criterion statements about basic skills and assessments of personal qualities are checked using boxes and achievements in individual subject areas are recorded using a norm-referenced four point scale.

4. Student self-records continuously available

A learning management scheme using negotiation between students and teachers of contracts for work and further negotiation of the assessment statements which will record and report that work, will produce quite different profiles from types 1-3. Except in Figure 4, objectives, frames, grids, ticks, grades and text have so far been solely teacher directed — even in Figure 2, only the contracts, not the profile report itself, involved students directly. (Figure 5 had a line assessing 'Self Awareness': one wonders how that can be calculated from the teacher's side of the desk!)

There are two famous instances of this sort of profile. One (really a pair) is the Swindon Record of Personal Achievement (RPA), begun in 1970 in England, and its successor, the Record of Personal Experience, Qualities and Qualifications (RPE), begun in 1974, in Devon. The other is the Schools Sixth Form Tertiary Entrance Certificate (STC), an alternative study structure to formal public examinations operating in some schools in Victoria, Australia since 1976. The RPA/RPE pair involved each student recording events, achievements and experiences with considerable flexibility. Each entry had to be attested by an adult, by way of verification. The whole turned out to be an idiosyncratic yet *highly reliable* and *valid* report on the years of school during which it was compiled.

The STC system makes self-reporting optional. There are negotiations of what is to be learned with selection of objectives and content; there is also negotiation in framing and wording the report of the assessment. A self-assessment statement in the record of achievement is permitted, and in some schools, actively encouraged.

5. School-long assessment portfolios: continuously available and updated

Types 5, 6 and 7 have been called 'macroinitiatives'. That indicates the depth and breadth (as well as the sheer bulk) of the report being offered. An Inner London Education Authority scheme gathers together a portfolio of (i) results in public examinations, (ii) graded tests, (iii) classroom tests, (iv) contributions from teachers, (v) pupils and (vi) parents. All this is testimony to the directions, achievement and progress of the individuals concerned. Such a document, if published, may well be a transcript of a school's whole record of the passage through its curriculum of a particular student, warts and all. Two offcuts from such a substantial bulk of material are described as Types 6 and 7.

6. Pre-transition summaries of achievement; available when the student leaves school

As students leave school they might take with them the summary of the full Type 5 portfolio profile, recording only the very latest (and hence most reliable and valid) of that vast mass of formative and summative statements. The profile will cover the whole range of curriculum experiences, but in a shorter form.

A notable example is the Oxford Certificate of Educational Achievement. Here is how it is described:

> The P-component will take the form of a personal record, compiled by the student in consultation with a teacher, which draws on the formative experiences articulated by the student and teachers in all curriculum areas and also such experiences beyond the formal curriculum. The G-component will be a detailed statement of what the student has achieved. These achievements will be defined by explicit criteria. The student will have progress recognised as it occurs and will be able to identify learning objectives and negotiate progress within the curriculum. The E-component will record all external examination results and shows how OCEA's three viewpoints, which are expressed in its three components, are each implicit in every subject.

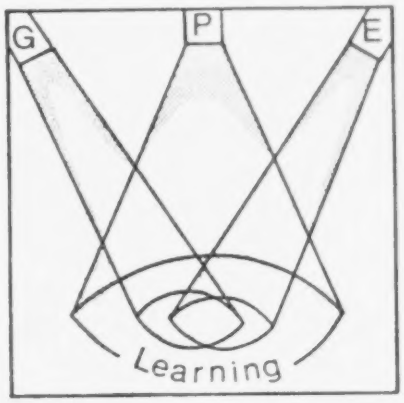


Figure 7: The P, G and E Components in the OCEA

Many people would not call it a profile at all. However, the OCEA has supporters:

> Most 'profiles' concern themselves with the products of education – reports, qualifications and certificates – rather than educational processes. The P-component however is about both process and product.

Another example comes from a school in the southern hemisphere. This school issues a reference certificate which is in effect a statement, in summary form, of teacher perceptions. Figure 8 is the student's application, Figure 9 is the report filled in by each of the three staff members 'who know you well'. The existence of such un-moderated inventories of pupil 'characteristics', in a computerised world of perpetual storage, is alarming, to say the least.

7. Vocationally-oriented achievement; reported when the student leaves school

A Type 6 summary, from a portfolio perhaps, when it provides only information which is relevant to the student's intended vocation or likely vocations, becomes Type 7.

I cannot find an exemplar. When I began to look around my large collection of profiles from five countries, I could not find one which was strictly vocationally oriented. The closest I came was Figure 3, already used for Type 2: this, at least, told the Rag Trade what Student X had done in a Clothing Construction course. But it didn't (and was not intended to) communicate details of curriculum, nor was its assessment fitted to the interests of the broad range of potential employers which Student X might have had in mind on leaving school. Admittedly some courses are directly vocational, such as the one which yielded Figure 5, but 'vocation' seems to be in the name of the Scheme, not the profile nor the capacities it represents.

APPLICATION FOR REFERENCE CERTIFICATE

(To be completed neatly by pupil at College Office. PLEASE PRINT IN BLOCK CAPITALS)

FULL NAME _____ Present Class _____
(Christian Names) (Surname)

Born: / / ; Entered Naenae College: / / ; Date Leaving: / /

Amount of Secondary Education: (years) (months)

Other Secondary Schools attended (with years):

Name THREE present staff members who know you well and would act as referees for you:

1 2 3

English Teacher: ; Maths Teacher (present or most recent)

Form Teacher: ; Tutor:

Principal Subjects Studied (N.B. only show the highest level for each subject, e.g. History F5);

English: F ; Mathematics: F ;

Fourth Form Option Subjects: Option 1 Option 2

Other subjects taken beyond Form 4 level:

to F	4 to F	7	to F
to F	5 to F	8	to F
to F	6 to F		

School Responsibilities (Prefect, Module Committee, Librarian, Team Captain, etc):

Forms 1 & 2 _____

Form 3 _____

Form 4 _____

Form 5 _____

Participation in Sport and Cultural Activities:

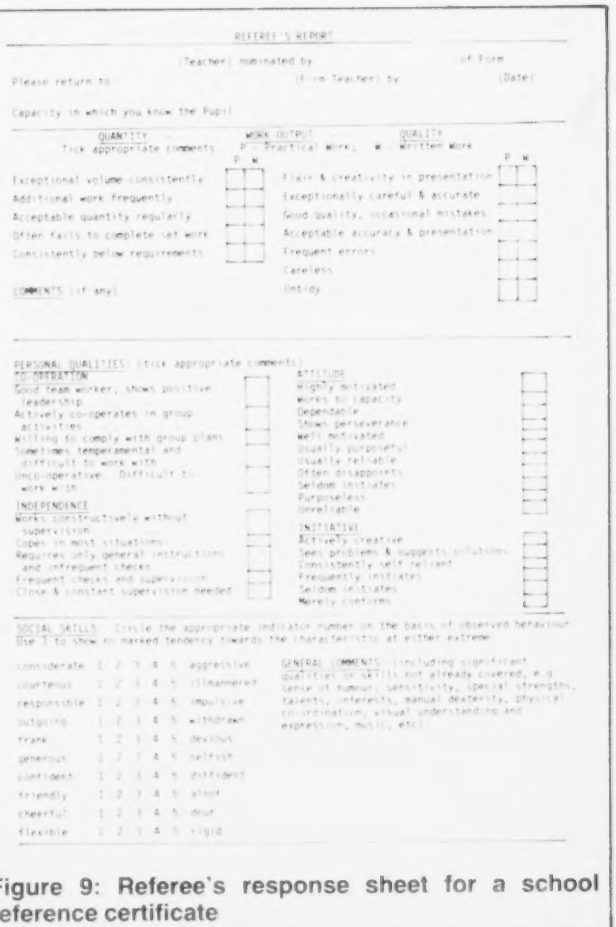
Activity or Team	Years	Team or Group This Year	Teacher Responsible This Year	Special Distinctions - major part in play, team captain, number of years in first team, awards, etc

Proposed Occupation:

The information supplied on this application is a true and correct account of my record at this school.

Date: Pupil's Signature: _____

Figure 8: Student application for a school reference certificate


REFEREE'S REPORT

(Teacher) nominated by of Form

Please return to (Form Teacher) by (Date)

Capacity in which you know the Pupil:

QUANTITY – Tick appropriate comments	P	W
Exceptional volume consistently		
Additional work frequently		
Acceptable quantity regularly		
Often fails to complete set work		
Consistently below requirements		

WORK OUTPUT P – Practical work, W – Written work

QUALITY	P	W
Flair & creativity in presentation		
Exceptionally careful & accurate		
Good quality, occasional mistakes		
Acceptable accuracy & presentation		
Frequent errors		
Careless		
Untidy		

COMMENTS (if any):

PERSONAL QUALITIES: (tick appropriate comments)

CO-OPERATION	ATTITUDE
Good team worker, shows positive leadership	Highly motivated
Actively co-operates in group activities	Works to capacity
Willing to comply with group plans	Dependable
Sometimes temperamental and difficult to work with	Shows perseverance
Uncooperative. Difficult to work with	Well motivated
	Usually purposeful
	Usually reliable
	Often disappoints
	Seldom initiates
	Purposeless
	Unreliable
INDEPENDENCE	
Works constructively without supervision	**INITIATIVE**
Copes in most situations	Actively creative
Requires only general instructions and infrequent checks	Sees problems & suggests solutions
Frequent checks and supervision	Consistently self reliant
Close & constant supervision needed	Frequently initiates
	Seldom initiates
	Merely conforms

SOCIAL SKILLS: Circle the appropriate indicator number on the basis of observed behaviour. Use 3 to show no marked tendency towards the characteristic at either extreme.

considerate	1	2	3	4	5	aggressive
courteous	1	2	3	4	5	clumsy
responsible	1	2	3	4	5	impulsive
outgoing	1	2	3	4	5	withdrawn
frank	1	2	3	4	5	devious
generous	1	2	3	4	5	selfish
confident	1	2	3	4	5	diffident
friendly	1	2	3	4	5	aloof
cheerful	1	2	3	4	5	dour
flexible	1	2	3	4	5	rigid

GENERAL COMMENTS: (including significant qualities or skills not already covered, e.g. sense of humour, sensitivity, special strengths, talents, interests, manual dexterity, physical coordination, visual understanding and expression, music, etc)

Figure 9: Referee's response sheet for a school reference certificate

Undoubtedly such profiles exist. But where? In the files of career guidance officers probably, the results of students taking such tests as 'career guidance inventories'. The profiles are unrelated, therefore, to the mainstream of what the students have studied and what has been assessed in the school. Like public exams, these guidance inventories are often detached from the mainstream of school life.

Eight features, generally agreed, which make up a 'good' profile

1. **A profile summarizes performance on a sequence of instruction.**

   It may incidentally offer 'diagnostic' insights, but this is not its prime function.

2. **It must report both academic and non-academic outcomes or achievements.**

   This feature is often questionable, and questioned by those who are shy of reporting 'effort', 'achievement' and other such attributes.

3. **It must have more than one point.**

   For 'point', one might read 'element', 'trait', 'skill', 'objective to be reported', or some such alternative.

4. **The 'points' are neither valid or reliable if added, or otherwise put together.**

   And this must be clearly indicated to the user.

5. **No letters or numbers are reported without an accompanying key.**

   The key will, ideally, be directly related to the content and process objectives of the learning.

6. **The objectives and format should ideally be recognised and used throughout the education system; the document should be part of a formal system of certification.**

7. **The complexity aimed at should shape the layout, etc., not just some administrative convenience.**

8. **The complexity aimed at should not compromise the reliability of the profile as a whole.**

Warnings

Warnings about the nature, content and limitations of profiles in general:

A simple checklist

For students, parents, prospective employers, guidance officers, university selection panels, and other tertiary educators.
Does the profile:

- ☐ name in full the student?
- ☐ carry a signature, stamp of authority, and date?
- ☐ record the year(s) of schooling to which it refers
- ☐ record the course(s) to which it refers?
- ☐ give a guide or key as to the meaning of any numbers (marks) or letters (grades) used?
- ☐ record the work that was done, in a way that any skills or achievements referred to can be related to something concrete or practical?
- ☐ indicate who the authors are (student/teacher(s)), and who the moderators are (other students, teachers, principal, etc.)?

Another checklist

This one is for all of the above, plus educational administrators, theoreticians, researchers, and politicians of any persuasion. Does the profile

- ☐ represent an intrusion on the personal privacy of the student?
- ☐ remain an unmoderated statement by one or a few person(s)?
- ☐ refer to a formative feed-back assessment, when it is to be used as a summative, marks only, statement?
- ☐ record a latest judgement which is more than two years old?
- ☐ record 'effort' and 'attitude' without details of how these matters were displayed in the classroom?

If it does any of these, discard those sections of it, and view the rest with suspicion.

Some important considerations

For anyone who designs, institutes, evaluates or is an audience for a profile.
How does the profile match up against the following opinions?

1. Profiles are progress reports, but there must be a place in any summative [marks only] profile for recognition of changes in the student since the last formative [feed-back] assessments were issued;
2. Profiles can report continuous assessment, but this must not be confused with, or replace totally, provision for a summative evaluation. Remember that an external unmoderated examination is not on its own an adequate summative assessment.
3. A profile should be more than just 'the disaggregated reporting of examination results'.
4. Profiles might purport to record 'mastery' but there are some fundamental implausibilities in such a concept – it is not generalisable across persons; nor continuously distributed across skills within one person; not generalisable from a base ability to a set of its component sub-abilities. What then is 'mastery' if it is recorded in a profile?

Profiles and Schools

Profile construction is an issue belonging firmly to schools. Here is an unedited and genuine letter, reprinted with permission, but anonymous by request, from an Antipodean parent to a principal about the institution of a continous assessment and profile

reporting system in his sons' school. It reveals in short form but with, I believe, superb clarity all the issues which might well disturb us.

Dear Jack

Third and Fourth-Form Reports

These are a few queries I have, and comments that I'd make off the top of my head, about the new form of reporting.

(i) There seems to be a definite move away from tangible, verifiable referents to more vague and subjective ones. Although the impression created is one of comprehensive information (many report forms and many headings) which is apparently quite precise (5-point scale of grades), I doubt if parents can take it all in and whether, in fact, the assessment is really as comprehensive and precise as it appears when the reports are subjected to close examination.

(ii) If pupils' progress is measured 'in relation to their own abilities', who assessed these abilities? How were they assessed (what measures, scales, systematic observation etc.), particularly the attitudes, effort, participation aspects which are heavily weighted in some subject areas to the exclusion of cognitive outcomes? Are the assessments of proven validity and reliability? When were the abilities assessed? (i.e., How recent is the information, particularly for 3rd Forms?) If these questions cannot be answered satisfactorily for *all* pupils on *all* criteria appearing in the report, then what follows in terms of the assignment of grades is likely to be highly inaccurate and misleading. In other words, if the teacher's assessment of abilities is inaccurate (either too high or too low) then damage will be done (pupils pressured to attain the unattainable; pupils achieving A grades without effort) and assessments will be awry because the baseline estimate was wrong. How confident can I be as a parent that the teachers in the various subject areas have *accurately* assessed my child's abilities, not only in cognitive areas, but in affective areas as well?

(iii) Presumably, a teacher having assessed a pupil's abilities will have some expectation about the progress such a child should make over the months covered by the report, i.e., a child of *n* ability should make *z* progress over *x* months (other things being equal?). The teacher is probably using internalised norms based on previous experience. But, these could vary markedly from teacher to teacher and such assessment would be extremely difficult for inexperienced teachers with limited exposure to representative samples of 3rd and 4th formers and particularly suspect in estimating the progress of very slow or very fast learners. What moderation is there of 'teacher expectancy' in assessing pupil progress? What does Teacher A expect in contrast to Teacher B both of whom are supposed to be assessing the *same* thing, but who are using *different* criteria (internalised, not stated, unexaminable) and who have had *different* teaching experiences. Is the Science teacher's notion of excellent progress for a bright, cooperative lad the same as that of the teacher of German? Is the A for Social Studies equivalent to a C for Mathematics? In other words, assessment is now entirely teacher-specific.

(iv) The Newsletter seems to indicate that progress is the result of 'hard work' on the part of the *pupil*, of building on established skills and developing weak areas. The onus appears to be entirely on the pupil to make progress in all facets of his school life. There is another component to the formula, of course. Suppose we have a pupil of good ability (however assessed) who is making mediocre or poor progress. Should the parent not ask the teacher why this is so? What responsibility has the *teacher* in ensuring that progress is commensurate with ability? If many pupils are making poor progress then maybe it is the teacher who should be looked at! Possibly his teaching is inefficient and/or ineffective, or pupils are insufficiently motivated and/or interested to achieve at an optimum level in his class. Presumably, if *all* teachers have matched their teaching styles with the pupils' learning styles, pitched their lessons at the right level, chosen appropriate and interesting material etc., etc., and provided excellent examples as adult models in terms of attitudes, etc., then *all* pupils should be shown to make progress over time. How likely is this? You and I know it's extremely unlikely, even if we discount that learning goes in fits and starts and that many of the outcomes are long term, particularly in trying to change attitudes and entrenched behaviours which may be deemed undesirable. So, what will a parent be anticipating in the second report? What should he expect if the teachers are doing their job?

(v) The aims stated at the top of the report forms are vague, woolly, typical of the kind which appear in most of our syllabuses. They are, in most cases, long-term ultimate goals which will not be achieved in the time span covered by the reports. Do you think they add anything for parents? If the teachers consider that aims should be included, why not have a statement of what was attempted *specifically* for the three terms covered by the report and how well each pupil achieved the short-term objectives?

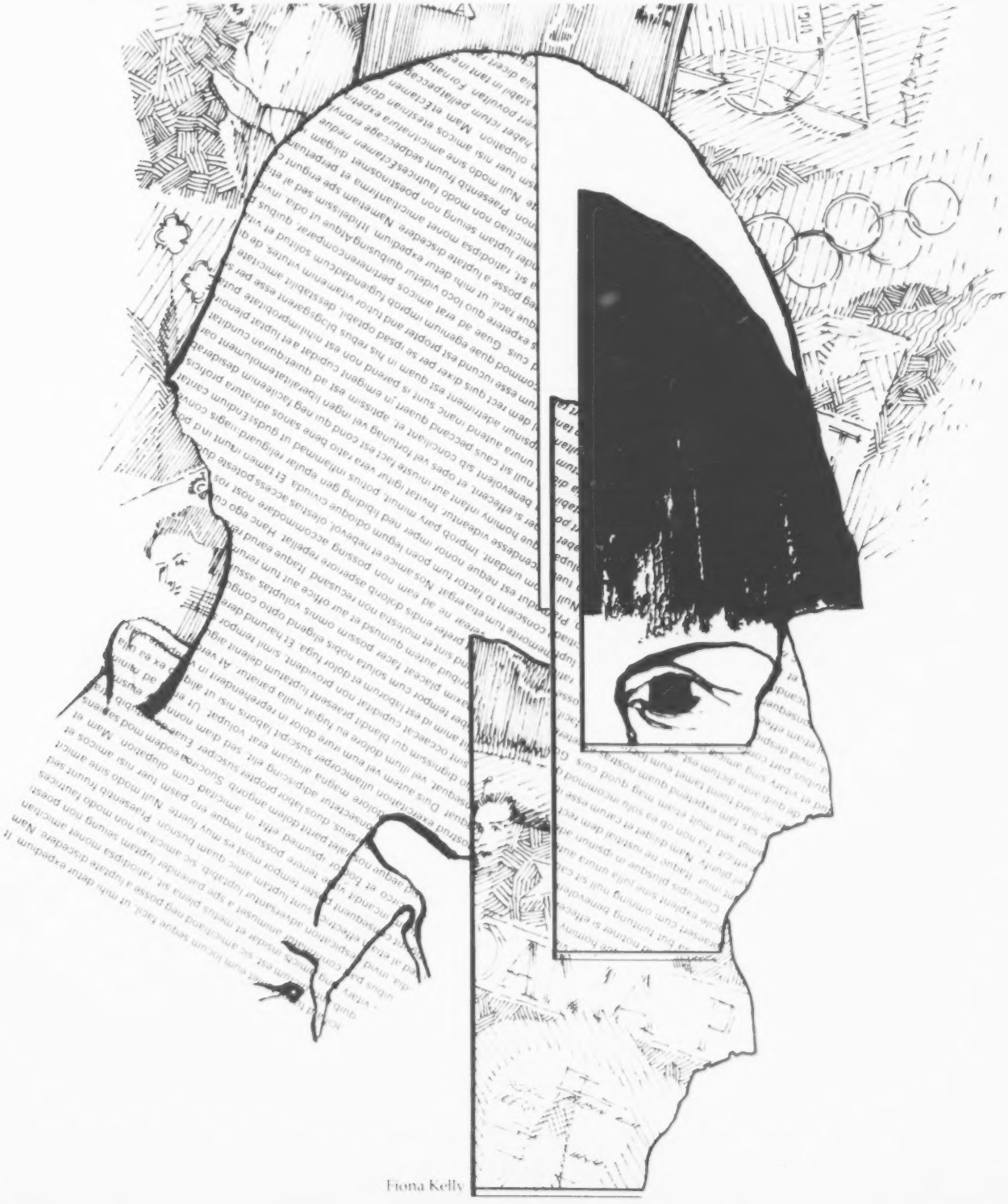
I could continue at some length, Jack, but I don't intend to, although I'd be happy to elaborate if you wish. While, as a parent, I applaud the school's efforts to provide me with a comprehensive picture of my children's progress I don't think I'm much the wiser, particularly in regard to academic achievement. And, as an ex-teacher, I'm sceptical about so many baldly-stated criteria which I know are extremely difficult to assess validly and reliably. Perhaps I'll be pleasantly surprised when I talk to subject teachers next week.

Kind regards,

Copying Permitted

© Copyright on this item is held by NZCER and ACER who grant to all people actively engaged in education the right to copy it in the interests of better teaching.

Non-Verbal Tests in Schools



Fiona Kelly

Non-Verbal Tests in Schools

By Cedric Croft
NZCER

Non-Verbal Tests

The term 'non-verbal test' is used to describe a range of paper-and-pencil tests designed to tap a selection of cognitive processes that are unlikely to involve verbal language. This does not mean that verbal instructions and verbal strategies have been entirely eliminated for all those who take these tests; it simply means that no words are included in the tests, the test content is of a non-verbal nature, and the responses to this content are unlikely to involve language.

Examples are the *Standard Progressive Matrices*, *Jenkins Non-Verbal*, *ACER Junior Non-Verbal*, the *IPAT Culture Fair Tests* and the *NFER Non-Verbal Tests*. All of these tests attempt to measure general intellectual skills of a non-verbal nature by utilizing shapes, patterns, diagrams and sequences. The use in schools of tests such as these is the focus of this article.

Performance and Apparatus Tests

There is a whole range of performance or apparatus tests widely used in psychological assessment which involve no verbal answers. The best known tests containing performance items are the *Wechsler Intelligence Scale for Children – Revised* (object assembly, block design) the *Revised Stanford Binet Form LM* (bead threading, paper cutting) and the *British Ability Scales* (block design, rotation of letter-like forms). The content of performance or apparatus tests is typically non-verbal, but the one-to-one administration of these tests, is a highly verbal process. It is the concrete apparatus that leads to performance tests being classified as a category that is separate and quite distinct from non-verbal paper and pencil measures.

Non-Language Tests

Some writers make a valid distinction between the true non-verbal test, and others that can be termed non-language, non-reading or pictorial tests. This distinction is useful, as non-language tests require no verbal language at all, and can be used for foreign-speaking, deaf and illiterate subjects. It can be misleading to call this sort of test non-verbal, since the term should only apply to the test *content* and not the subject's *behaviour*. Tests of verbal comprehension and vocabulary can be administered through the use of pictorial content, the outcome being that underlying verbal abilities are measured by tests with a content devoid of verbal material. Non-language tests are rarely used outside a cross-cultural setting, because in other settings examiners and subjects usually have some knowledge of a common language. The *Army Beta Tests*, previously used in large scale induction and recruiting in the United States, the range of tests used for selection in South African industry and the *Queensland Test*, are prominent examples of non-language tests.

The Characteristics of Non-Verbal Tests

Non-verbal tests:

(i) have a non-verbal content;
(ii) are in paper-and-pencil format;
(iii) have no apparatus;
(iv) are suitable for group administration;
(v) do not incorporate writing responses;
(vi) use oral language in administration;
(vii) tap cognitive processes unlikely to involve verbal language.

Why are Non-Verbal Tests Used?

The short answer is that non-verbal tests are seen as measures of 'ability' unconstrained by language, and because of this, it would seem that they can be used to measure cognitive functioning without being, as most tests are, dependent on language achievement. Non-verbal tests are thought to be of most value for testing children who are non-native speakers of English, or children whose measured verbal attainment is fairly minimal. But this takes a number of points for granted. What are the major assumptions underlying the uses of non-verbal tests in schools? Can these assumptions, and hence the uses of the tests, be justified?

Assumption 1: Non-verbal tests tap a set of thinking skills basic to all intellectual functioning, so they are measures of general intelligence.

It is wrong to assume that non-verbal tests measure the same cognitive functions as verbal tests, no matter how similar they appear to be. Spatial analogies are more than a non-verbal version of verbal analogies. The form of the relationships differ, the elements that lead to the perception of points of similarity differ and the level of thinking used to deduce the relationship is set at different levels.

Tests like the *Standard Progressive Matrices* and other similar non-verbal tests, have been designed to measure a broad selection of reasoning tasks and abstract conceptualization, but factor analytic studies have indicated that separate non-verbal factors are the greatest contributors to the scores. This suggests that the non-verbal abilities being sampled by these tests are largely distinct from the general verbal-educational (g,v:ed) factor being measured by verbal tests. However, it is probably misleading to think of verbal and non-verbal abilities as being entirely distinct. As verbal and non-verbal abilities are aspects of the broader group of skills now commonly referred to as scholastic aptitude (formerly 'intelligence', 'general intelligence', 'general ability', 'mental ability', 'general mental ability') it is probably more realistic to think of these abilities as two broad divisions of human intellect composed of a number of identifiable skills, with some general elements in common.

Studies of the relationships between measures of school achievement, scholastic aptitude tests and non-verbal tests, also shed some light on the relationships between non-verbal tests, and general intelligence. Typically, achievement tests sampling predominantly verbal skills (reading comprehension, vocabulary,study skills, spelling, writing skills) correlate more highly with measures of general intelligence (.80) than with non-verbal tests (.60). Mathematics involving problem solving also relates more highly to verbal than non-verbal tests, but aspects of mathematics stressing spatial skills, i.e. geometry, show a more positive relationship with non-verbal tests.

Assumption 1 cannot be supported. Non-verbal tests are not measures of general mental ability. Non-verbal tests are measures of the broad domain of non-verbal abilities, and cannot be used as valid measures of the intellectual skills associated with most of the highly verbal tasks commonly encountered in much classroom learning.

Assumption 2: Non-verbal tests are more valid measures of the school potential of the low-achiever than verbal tests.

Non-verbal tests are most certainly valid measures of non-verbal abilities, but this class of abilities is little utilized in most school learning. Non-verbal tests do not lack validity *per se*, but their validity is suspect when they are used to predict possible future achievement in verbal areas of the school curriculum. When the tests are used to predict verbal learning, there is a mismatch between the skills measured by the test and the abilities that underlie the learning. This cannot enhance test validity, as the tests are measuring something different from the skills and abilities underlying the intended achievement.

There is a widespread belief that poor readers who score well on a non-verbal test are likely to succeed in a remedial reading course as they have demonstrated a potential previously untapped by verbal measures. They may of course show very gratifying progress, but it is unlikely that such progress is due to the presence of the abilities measured by a non-verbal test. Reading is obviously a highly verbal set of skills and the abilities underlying this process are similar to those measured by verbal tests. There is little basis for the belief that non-verbal tests are satisfactory predictors of reading achievement, particularly of comprehension skills.

Where non-verbal test performance is very much better than verbal test performance it is tempting to infer that the non-verbal test is the more valid measure of underlying abilities, and that results from the verbal test represent a form of underachievement. It is then surmized that, given a different set of environmental circumstances, performance on a verbal test may have been nearer to the results of the non-verbal test. However, if you want to estimate an individual's present functioning and the likelihood of their progress in reading in the near future, a verbal test is best. It will tap present accomplishments in the skills that underlie reading. The hypothesis advanced was tempting, and widely differing scores may suggest a need for long term intervention, but for the purposes of planning for immediate short-term needs, verbal tests provide the more valid information.

There is one important exception to this general conclusion. In the case of students whose English is limited, there can be justification for using non-verbal tests. Verbal tests will only be valid if the subject has had considerable exposure to, and experience of, the language of the tests. If there was a need to undertake an assessment of a recently arrived child from Europe, Asia, or the Pacific, a child with little experience of English, a non-verbal test could give some very general notion of broad intellectual status. A verbal test in English would have little or no validity. In a situation such as this, it would be preferable to have any assessments made by an experienced psychologist who would have access to a range of valid tests.

Assumption 2 cannot be justified: Particularly in the short term, and provided the individual has knowledge and experience of the language medium being used, verbal tests are better predictors of most school achievement than existing non-verbal measures. Significantly better non-verbal scores may indicate cases of verbal underachievement, but the type of intervention required is beyond the resourses of most schools, and it is likely that the optimum stage for learning such skills is well past.

If non-verbal tests are to be used successfully in the way many teachers want, new tests with appropriate validating criteria must be constructed.

Assumption 3: Non-verbal tests are culture free.

There is a large body of evidence to suggest that non-verbal and performance tests may be *more* culturally biased than language tests. Cole and Hunter found the WISC Performance Scale to be as difficult as the Verbal Scale for a group of Negro children in the United States, despite the apparent cultural bias in many of the vocabulary, information and comprehension items, comprising the WISC Verbal Scale. Higgins and Sivers found that, when the *Revised Stanford Binet Form LM* and the *Coloured Progressive Matrices* were compared for a matched group of 7–9 year old Negroes and Caucasians, there were no significant differences in the Binet scores, but on the *Coloured Progressive Matrices* the Negro group did significantly *worse*. Vernon in 1965 reported that Jamaican boys scored better on conventional verbal intelligence and achievement tests, despite their linguistic disadvantage, than they did on non-verbal tests that appeared to be a 'purer' measure of general mental ability.

As non-verbal tests are apparently based on a white middle-class conception of 'logical thinking', should they be regarded as culture free? Cohen has suggested that there are two basic cognitive styles, analytic and relational, both regarded as independent of general mental ability, able to be defined without reference to specific content but, to some extent at least, influenced by social and cultural factors. If Cohen is right and the analytic cognitive style is implicit in non-verbal tests developed by white middle-class psychologists, the appropriateness of these tests for cultural groups that operate with a relational cognitive style, must be questioned.

Vernon has suggested that the group of skills we refer to as intelligence, or general mental ability, is bound up with convergent problem solving, persistence, initiative and efficiency. It is the type of ability well adapted to scientific analysis, control and exploration of the environment, large-

scale and long term planning and carrying out materialistic objectives. This has led to growth of complex social institutions (nations, armies, multi-national companies, school systems) but has been less successful in promoting solutions to group rivalries, or harmonious personal adjustment, than skills which would be called 'intelligence' by some cultures we regard as more primitive. If Vernon's suggestions are correct, and conceptions of intelligence differ from culture to culture, there must be little prospect of being able to use tests from one culture, as valid measures of the trait of intelligence, within another culture.

Major writers in the field of psychological testing agree that there is no such thing as a 'culture fair' or 'culture free' test, especially since there is no universal culture that test items can validly measure. 'Culture fairness' is not an either-or attribute but rather a number of dimensions along which various aspects of tests can range, and so is a matter of degree.

Assumption 3 cannot be justified. No test is culture free.

Do Non-Verbal Tests Measure Intelligence?

There is no simple, unequivocal answer to this question as it stands, as any conclusions depend very much on your definition of intelligence.

If intelligence is regarded as a broad grouping of cognitive skills manifested by the ability to see relationships, deduce similarities, solve problems, predict consequences, reason logically and generally manipulate a variety of symbols and the ideas they represent, non-verbal tests will be recognized as sampling *some* of these skills. Consequently, non-verbal tests will be viewed as a measure of some aspects of intelligence.

How do these tests relate to school learning? In the broadest sense, school learning is a result of interaction between the intelligence of the learner and the school curriculum. If we call the aspects of intelligence associated most closely with school learning 'scholastic aptitude', we must ask whether non-verbal tests are also measures of scholastic aptitude? It is unlikely that non-verbal abilities are a predominant aspect of the range of skills and abilities referred to as 'scholastic aptitude' because school learning is predominantly verbal.

To return to the question – do non-verbal tests measure intelligence? In the writer's view, non-verbal tests measure in the broad domain of intelligence, but when we are thinking of school learning we are most concerned with aspects of intelligence that can be categorized as 'scholastic aptitude'. While non-verbal tests are not comprehensive measures of scholastic aptitude, they cannot be disregarded entirely, particularly if it is accepted that non-verbal tests measure some aspects of intelligence.

What *are* the Legitimate Uses of Non-Verbal Tests?

Within the school context, non-verbal tests are most useful in guidance. They can be an aid in determining the range and strength of an individual's cognitive abilities, as a first step in career planning. Not every student who seeks guidance about future careers should be administered a non-verbal test. These tests should be among the resources available to a counsellor, along with interest inventories, tests of specific abilities (e.g., computer programmer aptitude, clerical and office skills, mechanical aptitude) tests of scholastic aptitude, measures of attitudes and personal adjustment. Non-verbal tests are no more, or no less important, than these other categories of tests. The key to their utility is knowing when they will provide valid measurements and when they can be used with profit.

In the case of a student with mediocre verbal achievement and significantly higher non-verbal test scores, a case could be made for this student to follow a secondary or tertiary course that utilizes non-verbal strengths, i.e., technical drawing, practical engineering or building. The associated difficulty is that although courses such as these utilize non-verbal skills, much of the associated instruction is undertaken by verbal means.

When you cannot use a verbal test, for example your pupil has recently arrived from overseas and you have no tests in his or her language, a non-verbal test may give you an approximate measure of intellectual status and possible achievement in the short term. It would be unwise to attach too much weight to these results, but in this situation non-verbal tests do enable a preliminary assessement of non-English speaking students to proceed.

These tests, like any other, provide specific information to be used in conjunction with information from a variety of sources. Provided non-verbal test scores are accepted as measures of non-verbal skills, they have a role to play. If they are thought of as a measure of a broad general ability that underlies school learning, they will be less useful. If they are regarded as being predictive of highly verbal school learning, they will be downright misleading.

Notes

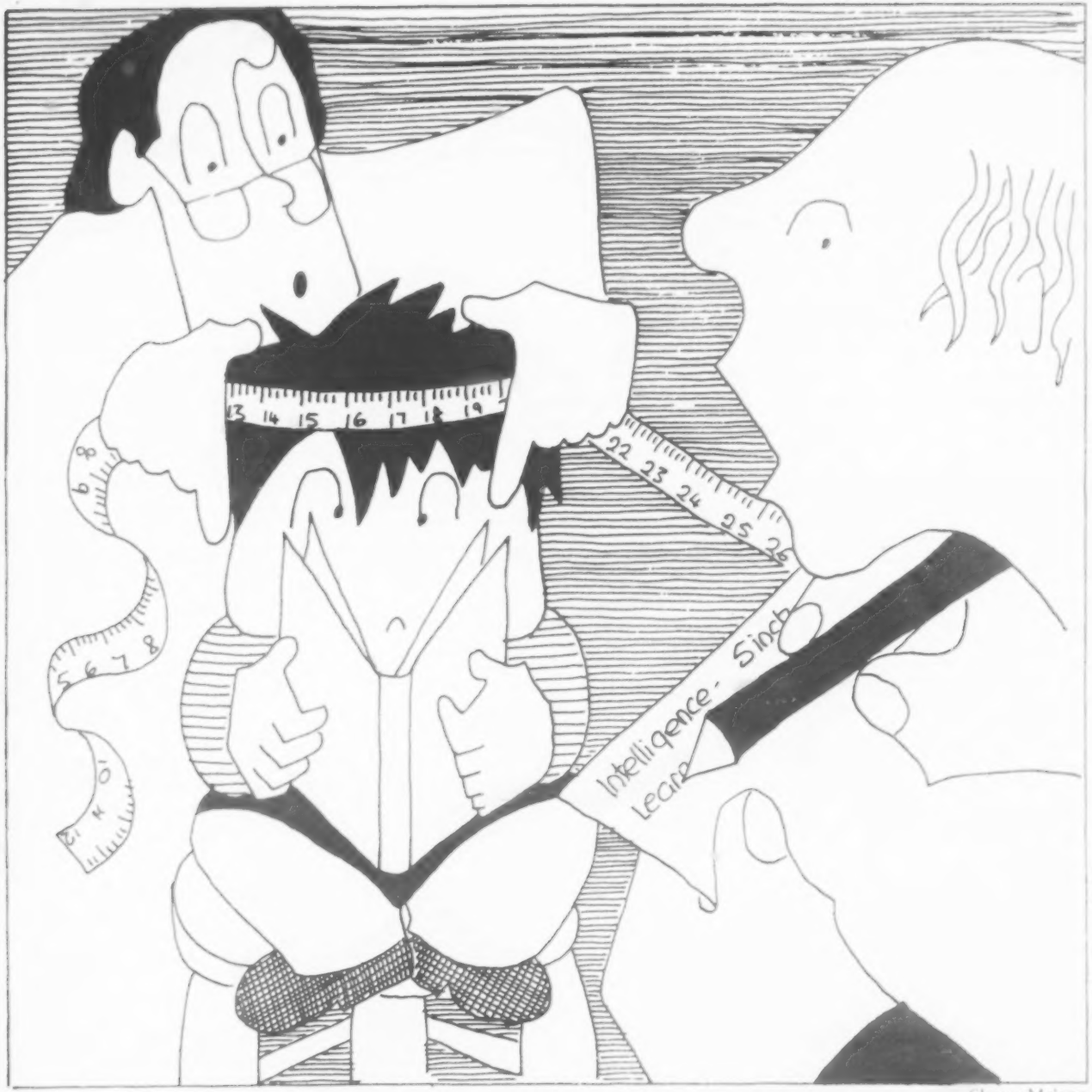
Evidence that non-verbal tests are not culture free can be found in Cole, S. and Hunter, M. 'Pattern Analysis of WISC Scores Achieved by Culturally Disadvantaged Children.' *Psychological Reports*, Vol. 29, pp.242–251, 1971.

Higgins, C. and Sivers, C. 'A Comparison of Stanford-Binet and Coloured Raven Progressive Matrices IQs for Children with Low Socioeconomic status.' *Journal of Consulting Psychology*, Vol.22, pp.465–468, 1958.

For discussion of why non-verbal tests are not culture free see
Cohen, R.A. 'Conceptual Styles, Culture Conflict and Nonverbal Tests of Intelligence.' *American Anthropologist*, Vol.71, pp.828–856, 1969.
Vernon, P.E. 'Ability Factors and Environmental Influences.' *American Psychologist*, Vol.20, pp.723–733, Sept. 1965.

Further reading

Sattler, Jerome. *Assessment of Childrens Intelligence*. Philadelphia: W.P. Saunders, 1974.
Anastasi, Anne. *Psychological Testing* (4th Edition) New York: Macmillan, 1976.



Shona Mclean

Does Intelligence Equal Learning Ability?

By Jo Jenkinson,
ACER

Introduction

As every teacher knows, in any normal classroom and whatever subject is being taught, there will be a wide range of individual differences in students' learning. Some will learn more quickly than others; some will retain more. Whether learning is measured by results on achievement tests, by the number of units completed, by time taken to progress through a set programme, or by the amount of material recalled at a later date, these differences will occur. Often it is the same students who repeatedly do better no matter what the subject, but sometimes surprises occur and a child who was thought to be somewhat dull will suddenly shine on one particular task.

So is there a single trait called learning ability, and if so, is it the same thing as intelligence? Do the more intelligent students always learn better than the less intelligent?

Some Definitions

Neither learning ability nor intelligence are clearly defined attributes of a person as are height and weight. Rather, they are concepts which the psychologist uses to explain performance on tasks demanding intellectual skills. But there the similarity ends. For a start, let's look at some common definitions.

Intelligence

However we seek to understand the nature or origin of intelligence, in practice it is used to explain variations between individuals on tasks requiring some sort of cognitive performance. Usually we obtain a measure of an

individual's intelligence by giving an intelligence test under standardised conditions. The test may use different types of questions, for example, vocabulary or number reasoning, to sample various cognitive skills at one particular time. The individual's level of intelligence is interpreted by working out his or her deviation or variation on the test from other individuals who are comparable in age or grade. But the tests do not tell us how the individual came to learn or acquire the skills sampled.

Intelligence tests are often called tests of learning ability because they are most commonly designed to distinguish between good and poor learners, especially in school learning. But they do not necessarily point to the underlying differences between students which can be seen when they tackle learning tasks. Traditional intelligence tests are simply a systematic, although very useful, way of comparing individuals' performances in order to make predictions about their likely efficiency in other intellectual tasks.

Learning

Learning, on the other hand, is a concept used to refer to changes in a person's performance during the course of practice, apart from those changes which are due to extraneous factors such as maturation, fatigue, changes in motivation, and so on. So, in principle, laws of learning could be derived from observations and experiments on a single person studied over a period of time.

In measuring intelligence we are interested in differences between individuals, whereas in measuring learning we are interested in differences within individuals before and after practice or instruction.

Can we then define intelligence as the result or product of learning? This would mean that a person's IQ, defined as the ratio of mental age to chronological age, would be an index of the rate of learning or information acquisition — it would show how much an individual has learned in a given time. But some forms of intelligent behaviour seem to require skills or concepts which cannot be taught until the individual has reached a certain stage of maturational development which allows more complex forms of learning. For example, we know from the well known experiments of Piaget that it is not until after the age of about seven that children can think according to the rules of formal logic, and then only if the task involves concrete objects that the child can see and handle. So intelligence appears to reflect maturational growth as well as past learning experience.

Thus learning and intelligence, although both hypothetical constructs used to explain cognitive performance, are derived from different types of measures, and for this reason they cannot be readily equated unless there is strong evidence of an empirical relationship.

What is 'Learning Ability'?

The first step in investigating the relation between learning and intelligence is to establish that there is in fact a single construct which can be termed 'learning ability', in which there are reliable individual differences which can be shown to be related to individual differences in intelligence.

If you give a class a set amount of work in a new topic, such as fractions, you will find after a time that some children have learned more than others even though none of the children had any knowledge of the topic to start with. Equal teaching does not mean equal learning. Much early research on individual differences in learning was influenced by the finding that an equal amount of learning experience, or practice, tends to increase differences between people rather than equalise their performance. It was concluded that the differences were due less to previous learning than to some sort of capacity for learning through practice. This capacity was presumed to be an innate characteristic and was termed *learning ability*. But results of subsequent investigations into this supposed capacity proved inconclusive. In one study, 56 students had practice in seven different intellectual-type learning tasks over a period of 39 days. Learning was measured by the amount of gain or improvement in scores over this period. Factor analysis of their gain scores on a wide variety of tasks revealed no general factor; instead, nine separate factors were found, and these involved rather limited categories of performance. Moreover these factors could not be interpreted as measures of learning but appeared to be more closely related to performance on fairly narrow tests of ability, such as memory, visual-spatial ability, speed, and perceptual ability. So there seemed to be no common, unitary factor of 'learning ability' which could form the basis of a relationship between learning and intelligence. Further, correlations between intelligence and amount of improvement were generally insignificant and often close to zero.

A further result discouraging any equation of learning ability and intelligence was that factor loadings for some of the learning tasks changed between the initial and final scores. For example, the loadings on a verbal comprehension factor decreased between initial and final trials on all seven tasks, suggesting that verbal comprehension ability influenced learning in early trials, but not in later ones.

In conclusion, the experimenter expressed doubts about the use of crude difference scores for the purpose of finding a common learning factor. Individuals do not always learn at a steady rate: the amount of increase in score with practice may vary. Usually we learn at a faster rate when new to the task. Figure 1, using a typical learning curve, shows the effect.

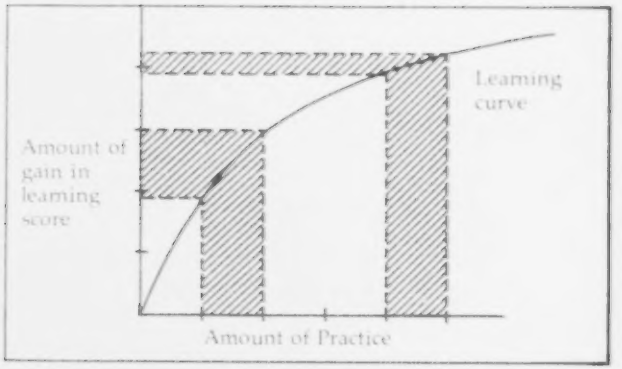


Fig. 1 *Typical learning curve showing amount of improvement at different stages of instruction.*

Many learning tasks have a 'ceiling effect': beyond a certain stage of acquisition people are unlikely to improve further, or the nature of the task does not allow further improvement. Further, changes in individual differences as a result of practice might be a function of changes in the way a task is performed during the course of practice. Anyone who has learned such a complex skill as driving a car will know that many actions which required a great deal of concentrated co-ordination in the early stages of learning become increasingly automatic with practice, allowing attention to be diverted to coping with more difficult aspects of the task such as driving in heavy traffic. So in studying learning, it is essential that a stable behavioural baseline, based on the individual's prior learning history, and relevant to the task at hand, be established.

To avoid some of these problems, psychologists have

attempted to devise tasks which may appear to be somewhat artificial, but are assumed to be relatively independent of previous learning so that everyone starts from the same base. Although the learning of such material as nonsense syllables may seem far removed from classroom learning, it seems likely that similar sorts of cognitive processes may be involved. At the same time the development of more complex statistical techniques has enabled experimenters to take into account variations in individuals' initial states of learning, as well as changes in the learning curve during the course of an experiment, thus improving the chances of obtaining more reliable measures of learning ability.

Studies of Learning and Intelligence

Employing some of these improved techniques, some studies in the 1960s attempted to establish a single factor of learning ability which could be related to intelligence. One study set out to investigate individual differences among children on twelve learning tasks, including word matching, maze learning, memory for words, listening comprehension, picture matching, and number-pattern memory. Learning curves were established for each task, based on average performance, and measures of learning were obtained by comparing each individual's performance with both the curvature of the average learning curve and the point at which this curve flattened out. A battery of reference tests, including Primary Mental Abilities tests, the Otis group intelligence test, and Stanford Achievement Tests, was given, and inter-correlations calculated for all measures. Both measures derived from the learning curve correlated positively with intelligence and achievement, supporting the belief that intelligence involves ability to learn. But, in a factor analysis, still no single learning factor appeared. Instead there were four separate factors: two memory factors, a numerical factor, and a concentration factor.

A further attempt to obtain a clearer picture of the relationship between learning and intelligence was made by Duncanson in 1965, using a smaller number of tasks in which he systematically varied the content. Three types of learning tasks, concept formation, paired associates (learning apparently unrelated pairs), and rote memory, were combined with three types of material, verbal, numerical, and figural, to give a total of nine tasks. These were given to 102 sixth grade children who were also given a group intelligence test, school achievement tests, and a selection of marker tests from the Kit of Reference Tests for Cognitive Factors. Learning measures for each individual were obtained from their learning curves.

Duncanson's findings can be summarised as follows. With the exception of concept formation, which appeared to form a factor on its own, learning was related to intelligence, scholastic achievement, and to the reference tests. The tasks were found to be, in general, related to abilities in a way that was appropriate, for example, tasks involving words were related to verbal tests, those involving numbers and figures were related to non-verbal ability, and paired-associates and rote memory tasks were related to a memory factor. But there were also learning factors which were independent of test scores: as well as concept formation, there were also separate verbal and non-verbal factors which were distinct from the corresponding ability factors. The most encouraging conclusion from this research was that, although no single general learning factor appeared, learning tasks could be reduced to a smaller number of factors, and some of these factors were related to measures of intelligence.

Several researchers have attempted to establish a relationship between intelligence and learning by comparing how well normal and retarded people perform various learning tasks. Zeaman and House reviewed the evidence from several studies. In six studies they found no evidence of a relationship between intelligence and acquisition of simple, classically conditioned responses — we all learn just as fast when the task is as simple as responding to a dinner bell. But out of eighteen studies relating intelligence to simple discriminative learning, for example, learning which of several bells means that dinner is ready, which one means the telephone is ringing, and so on, twelve reported positive results, with the more intelligent subjects learning more quickly. Nine studies reported no reliable differences in learning by people of various intelligence levels. The overlapping three reported both positive and negative results! In general, those studies which found a positive relationship between learning and intelligence covered a wider range of intelligence levels, so that any differences in learning would be more likely to show up. But the major difference between studies giving positive and negative results was in the difficulty of the learning tasks used. Six of the studies producing negative results used tasks which were either too easy or too hard. For example, some required only very simple two-choice discriminations, another used a seven-item, non-verbal, paired associated discrimination task which was too difficult for both normal and retarded subjects.

Over the range of studies of learning tasks, Zeaman and House concluded that at least a low positive correlation exists between intelligence and learning, provided a wide range of intelligence is sampled and tasks of intermediate difficulty are used. An examination of the learning curves of bright and dull subjects on visual discrimination problems suggested that the essential difference between the two groups was how long it took for improvement in performance to begin, rather than the rate of improvement once it started. This seemed to be associated with differences in attention between the two groups.

Towards a Unifying Theory of Learning and Intelligence

One of the problems with studies which have attempted to relate learning to intelligence is that they have lacked any theoretical framework to suggest which learning tasks should be selected for study. The review by Zeaman and House, leading to the conclusion that attention seems to be a significant factor in determining differences in learning performance between groups differing in intelligence, provides the beginnings of such a framework. Another useful theory assigns a central role in the development of intelligence to the individual's ability to transfer previous learning to learning in a new situation. But more research is needed into the common elements of learning and intelligence. Recent research has pointed to two possible directions. Firstly, we can look at traditional measures of performance in the areas of either learning or intelligence and attempt to characterise that performance using concepts traditionally applied to the other area. Secondly, we can hypothesise common processes underlying both learning and intelligence and look for empirical relationships to test our hypotheses.

In the first type of research, the simplest approach is to begin with intelligence test items and the possible reasons for success and failure on those items. Estes illustrated this approach by analysing four types of test items.

The first, digit span, requires successful recall of a string of digits after they have been read aloud. After examining

the task, Estes concluded that success did not simply require good associative memory, but that much longer strings could be recalled using a strategy of grouping the items into sub-groups or 'chunks' of approximately three digits each. A second item type, coding or digit-symbol substitution, seemed to require both good perceptual ability and good short-term memory, so that the subject could distinguish the correct symbol or code and hold it in memory long enough to reproduce it in the answer box. Success in vocabulary items, particularly where the subject was required to produce a definition, seemed to involve a number of skills which could not be clearly separated. Word naming, a sub-test of the Stanford Binet in which the subject has to produce as many words as he can in one minute, excluding counting or sentences, seemed to be related to the subject's ability to recall words in categories: those who attempted a simple chain association procedure did less well.

An almost infinite variety of intelligence test items can be characterised in this manner, suggesting processes or doing the test items and in various types of learning tasks.

The second type of research assumes that both intellectual and learning performances can be reduced to a common set of processes, strategies, or skills. The sorts of processes which are being investigated are span of apprehension, speed of information processing, rate of decay or loss of information, retention of information in correct sequence, and speed of retrieval of information from long term memory. The aim of looking for variation in these basic processes is not to achieve a better means of classifying people, but to understand what brings about specific kinds of competence or incompetence in intellectual tasks.

Some Conclusions

Can we then give an answer to our original question: Does intelligence equal learning ability? The not very satisfactory answer seems to be Yes and No. It is apparent that some types of learning show a higher correlation with intelligence than others; the more complex a learning task, the more likely it is to be related to intelligence. A summary of research findings suggests how these types of learning might be identified.

Firstly, learning is more highly correlated with intelligence when it is intentional and the task calls for conscious mental effort. Learning involving simple repetition or rote memory may even be negatively related to intelligence if processes are contrived to interfere with the rote learning. In one study, a group of gifted children took longer than an average group to learn a set of verbal concepts because they expected the task to be more complex, and wasted time testing out various hypotheses instead of using purely associative processes.

Secondly, learning is more closely related to intelligence when the material to be learned is hierarchical, in the sense that the learning of later elements depends upon the mastery of earlier elements. The relationship is also higher if the nature of the learning tasks permits transfer from a different but related past learning experience, and if the material to be learned is meaningful in the sense that it is related in some way to knowledge or experience already possessed by the learner. Learning the essential content of a prose passage is much more highly related to intelligence than is learning the serial order of a list of nonsense syllables.

The relation between learning and intelligence is also higher when learning is insightful — when it involves 'catching on' or 'getting the idea', or understanding a principle rather than merely acquiring information.

In addition, the learning task should be age-related and of moderate difficulty and complexity in order to be related to intelligence. Some things can be learned almost as easily by a young child as by an adult, while other forms of learning are facilitated by maturation or 'readiness'. If a task is too complex, students at all levels of intelligence may resort to simpler processes such as trial and error learning.

Finally, learning is more highly correlated with intelligence at an early stage of learning something new than it is later in the course of practice. Practice makes a task more automatic, and hence less demanding of conscious effort and attention.

In our present educational system, general measures of intelligence have been used to predict school achievement with a remarkable degree of success, suggesting that school learning probably fulfills many of these conditions. But with the limited state of knowledge at present of the fundamental processes involved in cognitive behaviour, it might be safer to regard traditional intelligence or aptitude tests as useful in predicting the outcomes of learning rather than the learning process itself. The most commonly used scholastic aptitude tests are designed to predict the products of learning in a particular setting. They are not designed to predict the ways in which different students learn best, to measure basic processes that underlie various kinds of learning, nor to assess the abilities needed to learn a new task.

Notes

What is 'Learning Ability'?

Conclusions about the supposed capacity called 'learning ability' were challenged by H. Woodrow in:

Woodrow, H. 'The Ability to Learn', *Psychological Review*, 53, 1946, pp. 147–58.

Studies of Learning Intelligence

Some investigations of the relationship between learning and intelligence are reported in:

Stake, R.E. 'Learning Parameters, Aptitudes, and Achievements', *Psychometric Monograph*, No. 9, 1961. Reported in Duncanson.

Duncanson, J.P. 'Learning and Measured Abilities', *Journal of Educational Psychology*, Vol. 57, 1966, pp. 220–9.

Zeaman, D., and House, B.J. 'The Relation of IQ and Learning', in Gagne, R.M. *Learning and Individual Differences*. Columbus, Ohio. Charles E. Merrill, 1967.

Towards a Unifying Theory of Learning and Intelligence

A theory of intelligence as ability to transfer learning was proposed by Ferguson in:

Ferguson, G.A. 'On Learning and Human Ability', *Canadian Journal of Psychology*, Vol. 8, 1954, pp. 95–112.

An investigation of the possible reasons for success or failure on test items was proposed by Estes as one way of studying the relationship between learning and intelligence in:

Estes, W.K. 'Learning Theory and Intelligence', *American Psychologist*, Vol. 29, 1974, pp. 740–9.

Some Conclusions

The conditions under which learning might be related to intelligence were identified by Jensen in:

Jensen, A.R. *The Nature of Intelligence and its Relation to Learning*. Melbourne, Fink Lecture, 1977. Reprinted in *Melbourne Studies in Education*, 1978.

TEST BIAS! TEST BIAS?

By Neil Reid and Alison Gilmore,
NZCER

Introduction

It is November 3 and Jonathon Tetley-Jones has just left the room where he has been sitting School Certificate mathematics. He is angry and frustrated. 'They asked questions about everything I didn't swot; there were topics in this we haven't even touched this year!' he moans to his friend Nigel. 'It's not right! The exam was unfair; I didn't get a chance to show what I know!' and he slams the exam paper into his schoolbag in despair.

Across the city at a large contributing school Mrs Carew is sitting at her desk. She has almost finished tabulating the results of a scholastic aptitude test administered to her class by the local intermediate school as part of a pre-entry testing battery. 'The Maori and Island kids are bottom of the heap again' she notes. 'They didn't have much of a show with all that difficult vocabulary, those wordy mathematics problems and tough verbal reasoning items, especially as some of them don't have much of a grasp of English.' She muses, 'Still, quite a few of the kids did O.K.; some pretty high scores too! Big difference between those middle-class children from Hillcrest and those from the other end of town. What you'd expect, I suppose, with the test biased in their favour and discriminating so unfairly against the Polynesian kids.' With a heavy sigh, she closes the mark register.

In the downtown part of the city, Deborah Pagent, a university graduate in geology, is being interviewed by a firm of management consultants. As part of a test battery she is required to take in applying for a particular job is a well-known measure of mechanical aptitude. Deborah is nonplussed. She has not even seen some of the mechanisms depicted, let alone had any opportunity to learn how they work. 'It's discrimination,' she thinks. 'Fine for men who've had experience with these things, but not for women – no way. Definitely unfair; I haven't a hope of doing well!'

These statements are fairly typical of those that can be heard every day in and around schools or other places where tests are given. Are these legitimate complaints against tests? Are tests, as is claimed, really so biased and unfair to minorities, or to certain sub-groups? Unfortunately, it is impossible to give a straight 'Yes' or 'No' answer to these questions. Things are not as simple as they may appear. The whole topic of test bias is complex and confusing, partly because there are so many definitions of 'bias' (including common, everyday use), partly because the issues become highly emotional for those who see themselves as disadvantaged or who are championing a cause, and, to some extent, partly because of confusion in the interpretation, or meaning, of test scores.

The illustrative statements introducing this article touch upon some of the major issues. While there is no universally accepted definition of test bias, five broad types of bias may be identified: content bias, bias in language, atmosphere bias, bias in prediction

and selection, and bias in interpretation with social consequences.

The intention of this article is to clarify the issues so that a more enlightened debate can take place. Accordingly, we introduce these different kinds of test bias, examine the basis for the claims made against tests, and discuss whether such claims are justified. Lack of space precludes extended discussion, but several readily available references are provided at the end of the article for further reading.

1. Content Bias

Most claims of bias in a test concern its content. Each of the three illustrations contains elements of content bias, which is the type of test bias that comes readily to mind for most people.

A test with biased content contains questions that in one sense or another are 'unfair' to an identifiable sub-group of those being tested. Examples of typical bias cited include: words that are unknown or unfamiliar; topics that have not been covered adequately or at all; questions drawing on experiences outside the range of those normally expected for an age, class or minority group; emphasis in the tests on values and skills that are traditionally middle-class European, and so on. The concerns are genuine. Are they justified?

A test is normally given, in the first place, to gauge how many of the questions it contains can be answered correctly. What is sought is information about the ability being assessed, as revealed by the answers to the questions – *not* reasons for inability to achieve, such as restricted opportunity to learn the content, or failure to take advantage of that opportunity. Unquestionably, middle-class European children in our society are better fitted to take these tests. Their more extensive experience means that they have had a greater chance of acquiring the kinds of skills and competencies which underlie school performance and which are incorporated in scholastic-type tests emphasizing verbal and numerical abilities. But this does not make the tests biased against other students. As long as the tests can accurately assess present attainment and predict the school performance of children, it makes no difference for the *validity* of the tests how the children have come by the skills and knowledge. The tests still serve the purposes for which they have been designed.

Turnbull, speaking before a U.S. Senate Subcommittee, in 1977, saw it this way:

> The test score tells you how well the student has mastered the skill in question. It does not tell you the obstacles he or she has overcome to attain the degree of proficiency. If one is concerned with helping students develop a level of skill necessary to get along in our complex society, it is important to be able to measure attainment separately from the question of how the learning was or was not acquired.

Eliminating Content Bias

Widely used published tests are usually carefully constructed and, in the case of achievement tests, they sample content that is specified by an accepted syllabus, curriculum, textbook series and so on. In this type of test, content validity is of paramount importance. Aptitude tests also sample a domain that is described and delineated by the test maker. In both cases the tests are subjected to 'sensitivity' reviews, usually involving members of identifiable minority groups, in an effort to detect and eliminate content or questions that are offensive, unfair, ambiguous or inappropriate. Words, phrases or descriptions considered in any way biased are removed.

In addition, the tests have to meet rigorous statistical checks, and trialling is done with the kinds of students who will eventually take them, including those from cultures and backgrounds outside the mainstream. Where there is evidence of particular sub-groups performing differentially on specific items, such items are carefully scrutinized for possible content bias. If a source of bias is detected, the items are either re-written to eliminate the bias or discarded.

While it is possible to compose 'bad' tests, reputable test developers try to meet stringent professional standards and requirements, including several that apply to test bias (see, for example, *Standards for Educational and Psychological Tests*, 1974). Test makers must be constantly alert to possible sources of bias in their products. Even the appearance of discrimination needs to be carefully avoided!

Achievement and Aptitude

Before proceeding, we will make an important distinction between achievement and aptitude tests and the interpretation of scores made on them. It is a difference that arises again and again and is crucial to a clear understanding of the bias issues.

Despite a prevailing myth to the contrary, tests of pure aptitude do *not* exist. Since current aptitude cannot be measured in isolation from past achievement, the two types of test obviously overlap in content. However, there *is* a great deal of difference in how the results of such tests are interpreted. And it is this aspect, with its considerable social ramifications, that triggers the emotional reaction attached to test score interpretation.

A high score on an *achievement* test can be interpreted legitimately as an indication that the content of the test has been learnt and that further instruction in the area examined is probably not necessary. A low score on the same test may be interpreted as evidence of a lack of attainment and hence further educational effort or input is required.

On the other hand, an *aptitude* test, instead of measuring attainment, is used in predicting future achievement on the basis of present accomplishment. A high score will be interpreted as evidence of ability to learn now and in the immediate future. But a low score may be interpreted as showing that the pupil has insufficient capability to achieve and in some cases certain educational resources and opportunities to learn may therefore be withdrawn. Hence, in the difference between the interpretation of achievement tests and aptitude tests lies a decision of considerable social significance. In the former, a low score prompts an increase in education resources to boost achievement; in the latter, the same low test score may result in resources being withdrawn or denied – the very resources that might help to alleviate the lack of abilities or competencies. However, to assign pupils to slow-learning classes on the basis of a test score alone, to water down the curriculum and to lower expectations on their behalf, will do nothing to enhance the learning chances of low-scoring students.

Failure to make this distinction between achievement and aptitude can lead to claims of bias. Where Maori students as a group, for example, score low on the *PAT: Reading Vocabulary Tests*, the charge is one of 'bias' and may lead to an abandonment of the tests, rather than a demand for an improvement in the educational system. Yet when the tests are subjected to close scrutiny, both judgemental and empirical, the test content is found to be a sample of words taken from a whole range of reading material to which children are exposed both inside and outside school – exactly as one might expect. The test scores do not indicate a lack of capacity to achieve. Instead, they provide evidence that additional educational effort, perhaps more of the same, perhaps a different approach, is required since achievement is lacking. If you have a high fever you don't solve the problem by smashing the thermometer. The challenge is surely to use the information as an aid for identifying the most effective education experiences for the pupils.

Achievement Elsewhere?

Minority groups sometimes argue that their children *are* achieving in the school system, and it's just that the tests are asking the wrong questions and testing the wrong things. The implication, of course, is that somewhere there is a group of 'right' questions which, if posed, would show satisfactory achievement by the minority students. And, conversely, we often have the same people claiming the system has failed their children, citing the low achievement, as indicated by test scores, to prove the point! Such a

claim is justifiable only if the test questions are invalid and do not represent a balanced sample of curriculum content, what has been taught, textbook content and so on.

Mean Score Differences

It is often claimed that pupils from ethnic minorities or low socio-economic status homes achieve lower scores on achievement and aptitude tests, because the tests are biased. This view is mistaken; mean differences *in themselves* are not a legitimate standard for identifying bias. If they were, as Gardner says, for achievement tests, 'every spelling test would be biased against poor spellers, every vocabulary test against persons who had poor vocabularies . . .'

In considering why such score differences are revealed by tests, Flaugher, in his cogent 1978 article on test bias, states the position succinctly:

> . . . it would be surprising if most kinds of tests didn't show mean differences in favour of the majority group. It would have to be a peculiar kind of test indeed to fail to reflect the disparities and differing advantages that are so evident through other sources. Yet many critics of testing merge the concept of equality of *opportunity*, which is certainly a legitimate goal to be sought, with the concept of equality of *result*; but it is only results that the tests in fact measure. The existing discrepancy is evidence that the legitimate goal has not been attained; to accept the discrepancy instead as evidence of test bias is to deflect attention from the pursuit of that legitimate goal.

Clearly it is the opportunity to learn which is crucial. Since the test maker must work on the basic assumption that all students have had reasonably similar experiences and backgrounds before taking the test, where students have been denied the opportunity to learn because of restricted background and/or disadvantaged home and/or school environments, the group differences will often simply reflect this rather than true differences in ability. Knowledge of the test taker's circumstances must always be taken into account in deciding on a test's suitability and in interpreting the test scores (see TOSCA: Teacher's Manual, p.4).

Fairness in Relation to Use

To complicate the issue, test fairness in terms of content can be evaluated only in relation to *use*. We are interested in assessing students' present performance on the tasks making up the test – tasks that are demonstrably appropriate and significant (i.e., valid for the intended purpose), rather than the sources of variance in the test scores (i.e., the reasons why some students do not achieve). Doubtless, those minority and disadvantaged students, whose experience and background have not equipped them with the competencies essential for school learning, will score low on achievement and scholastic aptitude tests. Insofar as these factors affect performance, their influence will and should be detected by the tests. However, those low scores will represent not a bias in the tests, but a genuine deficiency in those pupils – a lack of the particular kind of ability being measured. The tests tend to be accurately descriptive of current accomplishments. It is vital that unjustifiable and damaging inferences are not made about the innate capacity of an entire ethnic group or students from a particular SES category on the basis of the low test scores, or about the permanency of such ability deficits.

A further complication is that a content-biased test may be a valid and useful test in particular circumstances. Tongan learner-drivers, for example, might not have had the same opportunity to learn the English language as have other New Zealand students. So, a driving test that required the reading of road signs written in English, such as LOW OVERHEAD CLEARANCE, BRIDGE UNDER REPAIR, NEW SEAL LOOSE CHIPS, METAL SURFACE AHEAD SLOW DOWN, and so on, would be entirely appropriate. The test is obviously necessary, even though a sub-group of those taking it might be disadvantaged. Content bias is inevitably bound up with test use.

Limitations of Tests

While it is true to say that achievement and aptitude tests do a tolerably fair job of assessing certain facets of the vast domain of human behaviour, it might also be claimed that this limited spectrum has tended to become over-valued when compared with those other traits and abilities that are measured far less well, such as creativity, critical thinking, and valuing. As one writer has said: 'Since we can't measure all of the important things, we consider what we can measure all-important' (Flaugher, 1978). And when tests are seen to be tapping these 'valued' skills and abilities to the exclusion of others, the issue of bias is quite rightly raised by those who detect the deficiencies.

There is then, a discrepancy between what the test makers claim they can measure validly and reliably and what the public at large believes the state of the art of educational and psychological measurement to be. There is a tendency towards over-interpretation, and some members of the assessment enterprise must take some of the blame for this distortion. Their past claims for what they are able to measure have not always been conservative. It is something of a leap to go from failing to correctly solve several verbal and quantitative problems in a pencil-and-paper 'intelligence' test, to being labelled as 'unintelligent'. This kind of over-interpretation of the intent and the outcomes of testing is a legitimate aspect of test bias. Obviously, great care must be taken, by both test makers and test users, to evaluate test performance within the constraints of the test's content.

2. Language Bias

Whereas achievement and scholastic aptitude test results may provide a reasonably accurate indication of a pupil's current ability to deal with curricular materials and to cope with the knowledge, skills and understandings that predominate in school learning, what is presented in tests of the kind most commonly used will be written in standard English. In this sense they clearly favour those whose mother tongue is English. And, in this respect, such tests will place at a disadvantage those whose mastery of the English language, for whatever reason, is less than that of their age and class peers. But, it does not follow that the test is biased against these pupils, and that they would do better if the test was translated into their first language. It simply means that these students have poor English language skills, something that can be remedied. Nor does it mean that the tests should be discarded or that they are less useful for the purpose for which they were designed. Above all, it does *not* mean that a child who has had limited exposure to English and who attains a low score is inferior or stupid. Such inferences are unjustified and reprehensible.

Considerable responsibility devolves on the test user in instances of this kind. For example, why on earth give a *PAT: Reading Comprehension Test* to a Vietnamese student who has been in the country for three months? Such a use is obviously inappropriate and the test score will reveal nothing not known already.

Appropriate Language Levels

Test constructors must take care not to confound what they are seeking to measure by presenting their tests in vocabulary and syntax well beyond the level of the pupils for whom the tests are designed. If such care is not taken, the test would indeed be unsuitable for the poor readers. A distinction must be made here between the reading difficulty in, say, a mathematics test, of its mathematical terminology (which is entirely appropriate) and of other words appearing in the test stimulus material and questions. Essentially, the score on the test should reflect accurately the mathematics knowledge and understanding of the test taker and *not* his or her general reading ability. The bias introduced by the unnecessarily complex language will have a confounding effect on the measurement of the mathematical skills. The way round this is to make the measure more appropriate by modifying the language of the test.

Sexism and Sex Stereotyping

As Flaugher points out, there are great similarities in the unfair treatment of ethnic minorities and women. More than any other group, women are up against the English language itself: the language has a distinct masculine bias, particularly in the generic use of male nouns and pronouns when the content refers to both sexes. And this bias is reflected in tests as in textbooks, curricular materials and the like. In other words, if the curriculum, printed materials, illustrative examples and so on are slanted to any extent, or if they present current stereotypes, as they are bound to, then these same biases and cultural content will be reflected in the tests; especially if the tests are to sample content and curriculum emphases representatively and accurately – as achievement tests must.

It is up to the test makers to do their utmost to ensure that their products do not serve to perpetuate the image of sexual inequity or use sexist language when it can be avoided. As has been stated above, tests cannot lead this change, they must reflect the status quo, particularly in terms of content and emphases in achievement testing. What can be done, though, is to make others more aware of and sensitive to these subtle biases and to eliminate language and stereotypes offensive to women.

3. Test Atmosphere Bias

Claims are sometimes made that typical achievement and aptitude tests underestimate the actual performance of minority groups, because the test takers react negatively to certain aspects of the testing situation. Children from these groups, it is said, obtain low scores, not because they lack the abilities measured by the tests, but because of non-cognitive factors, as discussed below.

Interaction Effects

One of the factors that has received considerable attention is the interaction that occurs between the tester and the person being tested. Test administrators assume that those taking the tests will be motivated to do their best, and that by establishing rapport and using standardized administrative procedures 'tester effects' will be minimized, particularly for individualized testing, where the level of interaction is high.

The critics charge that, while these assumptions may hold true for middle-class European children, they usually don't hold true for those from other sub-cultures or socio-economic levels. To remedy this situation and to counteract likely effects, they have suggested that the tester should always be from the same ethnic, socio-economic or sex sub-group as those being tested. Apart from the impracticality of such a proposal, research into atmosphere bias provides no empirical evidence to clearly support the critics' contentions. Nevertheless, it would be unwise to ignore these potential sources of bias and test users should be alert to the possibility of such effects.

The Test Situation

Perhaps the concern expressed in this aspect of test bias would be better directed towards the whole social psychology of the testing situation? Maybe it is the very act of testing itself which is unfair for some persons in that they are unable to demonstrate their real capabilities. They might well be inhibited when confronted by a test because of past negative, hurtful experiences, which could be commonplace for certain minority groups.

As test users, we should perhaps ask ourselves whether we should administer nationally normed standardized tests to groups of students who, in terms of academic achievement, differ markedly from the majority of children. Perhaps other kinds of testing – tailored and/or criterion-referenced – might be more appropriate and less stressful? Of course, if one of the purposes of the school in using standardized tests is to relate the group's achievement to that of age and class peers nation-wide, then the testing of these students with appropriate standardized tests must be conducted; there is no other way to gain such information. However, it does seem counter-productive to inflict periodically both these students and their teachers with detailed evidence of just how far from the national norm they in fact are, and possibly snuff out any flickering enthusiasm for learning as well as adversely affecting teacher morale. As Flaugher has stated:

> . . students and teachers in [this] setting know this kind of test bias in a very personal and painful way and understandably are hostile, ready to condemn that process and testing itself as a demonstrably harmful influence on their lives.

In such situations an atmosphere of discouragement and disincentive is created by the assessment process itself, rather than by those potential sources of atmosphere bias identified earlier. With this in mind, test users should try to choose or to develop an assessment technique that will minimize these effects and strive to provide testing conditions consistent with the purposes of the assessment. Easily said – *not* so easily done!

4. Test Bias in Prediction and Selection

In a society like ours, in which equality of opportunity is universally accepted, the problem of bias in prediction and selection, whether for admission to educational institutions or for employment, is a troublesome one. Currently, tests are used extensively to predict future achievement and performance, to control entry to tertiary and specialist education, and to help in deciding about suitability or otherwise for particular vocations.

A test is considered to be biased, if the predictions or decisions based on the test scores vary for different groups. Matters of differential validity and predictive validity arise, and the discussion necessarily involves statistical considerations. Again, the treatment of this aspect of the topic must be over-simplified, but essentially the nub of the issue is as follows.

If a test is examined for possible differences in predictive validity for ethnic, sex or other identifiable sub-samples of the population for which the test has been designed, and if no significant differences are found between the groups, then the same decision-making rules can be used for everyone, regardless of group membership. The statistical technique employed is most commonly the calculation of the relationship between the test score and some criterion measure, such as School Certificate marks (4-best subjects), first-year university performance, an interview rating, or the number and value of sales made over a specified period of time. If the test is predicting equally well for each group, then there is no problem. But, if it is not, and the testing procedure is deemed to be invalid for one of the groups, then alternative assessment methods need to be devised for that group.

Much overseas research comparing prediction for whites and blacks, largely from the U.S.A., has concluded that differential validity is nonexistent, and that even if it does 'exist' it is not a very potent phenomenon; it is pseudo-problem. Contrary to what is often supposed, several of these investigations have revealed that test scores predict in favour of blacks, and that it is persons from the dominant culture, the majority group, who are being discriminated against! Paradoxically, the tests are not considered biased in this case. As Silverman has observed: 'Tests are only thought of as biased when they assign comparatively low scores to easily identifiable sub-groups.'

Biased Criteria

However, there are complications yet again. Some critics of testing have raised a noteworthy point. They contend that a test may be demonstrated to be valid for predicting, say, school success, but it may still be unfair to minority or disadvantaged students, because the criterion used in the validation study, for example, a score on the *Test of Scholastic Abilities (TOSCA)* (the predictor) correlated with 4-best School Certificate subject marks (the criterion), is itself biased. And that brings us back to the issue of the appropriate uses

to which the test results are put, and to the social consequences of our actions as test users. This aspect of test bias has received scant attention in the research literature, as Gulliksen has pointed out, and it presents a complex problem that is far from being solved.

What's to be done when the reliabilities of the predictor and the criterion are different for different groups? It has been suggested that this in itself causes the mean score differences so often observed between majority and minority groups. Queries about equivalence naturally arise.

What do we do when the criterion is biased for a minority group, yet it is traditional, well-tried, accepted by the vast majority, and cannot be replaced by anything remotely as good and useful? Just how sound (and unbiased) are our traditional criteria of external examination marks, supervisory ratings, internally assessed teacher grades, speed of work, and so on?

In the United States, court battles are being fought regularly over these and similar criteria and their use in prediction and selection. Legal action has been taken to challenge alleged discriminatory practices operating in selection for admission to tertiary educational institutions, in employment and most recently in school systems requiring 'minimal competencies' of their students before graduation. The result has been the formulation of a series of guidelines and regulations in an effort to ensure fair treatment for all.

Test makers and test users need to be alert to differences in validity for various identifiable sub-groups within our heterogeneous, multicultural population, and to modify their decision-making procedures accordingly.

Avoidance of 'Double Standards'

In matters of selection fairness, what we should guard against is any kind of 'double standard'. Take the situation where two candidates are up for selection. They gain *identical* scores on the prediction test, but are treated differently according to ethnic identity, with a *lower* requirement being accepted for the minority candidate. While any 'double standard' of this kind threatens seriously our treasured principle of 'equality of opportunity', it might be noted that we now have the concept of 'positive discrimination' entering into recent legislation, as in the New Zealand Human Rights Commission (1977) and Race Relations (1971) Acts.

5. Test Bias as Inappropriate Use with Social Consequences

Reschley has neatly summarized this important aspect of test bias. He states:

> The ultimate criteria that should guide our evaluation of test bias are the implications and outcomes of test use for individuals. Succinctly stated, test use is fair if the results are more effective interventions leading to improved competencies and expanded opportunities for individuals. Test use is unfair if opportunities are diminished or if individuals are exposed to ineffective intervention as a result of tests.

A classic example of an *unfair* use of test results would be the assignment of a child to a slow-learner class on the basis of a single test score, unsupported by other evidence. It was just this kind of abuse that resulted in a Californian court forbidding San Francisco School District psychologists the use of individual intelligence tests for the assignment of minority students to classes for the educable mentally retarded (Case of Larry P. et al vs. Wilson Riles et al, 1972). While Reschley takes a broad view that encompasses additional issues, fair uses of tests are clearly those which foster individual development, whereas unfair ones hinder that development. Simple and direct!

Darlington, taking a slightly different tack, reminds us that a test is a tool, and as such is not a bad device *per se*. '. . . it is the particular use of a test, not the test itself, which is fair or unfair.' The burden of responsibility shifts from the test maker to the test user whose duty it is to be aware of and to eliminate discriminatory circumstances.

Conclusion

The word 'bias' itself is almost always used in a pejorative sense; it evokes, as Gardner notes, '. . . affective, even visceral, reaction.' As educators who employ tests in our work we need to be absolutely sure how we're defining and using the word 'bias'. We should avoid using it loosely, and be mindful of its various aspects in test use and the interpretation of test scores.

Throughout the discussion the need to be alert to possible sources of bias in tests has been reiterated. These cautions must be taken seriously by all educators involved in testing on professional, ethical and legal grounds. But, equally important, as the APA *Standards* warn, is to avoid seeing bias where none is present. As test users, let us avoid chasing elusive and possibly imaginary bogeymen.

The issues are complex and confused. It is also doubtful whether all the factors that lead to bias have yet been identified, much less understood, controlled for, or corrected. Although psychometric techniques are undoubtedly improving, there is far from universal agreement as to their mathematical elegance, application and usefulness. And a purely technical resolution of the many problems of test bias is unlikely to be adquate as value judgements, ethical considerations and legal aspects are involved right down the line. In the long run it is far more likely that these matters, to paraphrase Mercer, will be settled in the political arena rather than in the halls of academe, regrettable as that may be.

On the other hand, while we have indicated that the testing fraternity is striving to grapple with the contentious issues, we should be mindful of the alternatives to achievement and aptitude testing. Robert Ebel comments on what some of the social consequences of *not* testing might be,

> . . . Excellence in programs of education would become less tangible as a goal and less demonstrable as an attainment. Education opportunities would be extended less on the basis of aptitude and merit and more on the basis of ancestry and influence; social class barriers would become less permeable. Decisions on important issues of curriculum and method would be made less on the basis of solid evidence and more on the basis of prejudice or caprice.

All things considered, maybe we shouldn't burn the barn to catch the mouse?

References

The statements by W.W. Turnbull, formerly president of ETS, can be found in

Turnbull, W.W. (1977) *Statement Before the Subcommittee of Education, Arts and Humanities Committee on Human Resources, U.S. Senate.*

Professional test makers' standards and requirements are fully discussed in

American Psychological Association, American Educational Research Association and National Council on Measurement in Education (1974) *Standards for Educational and Psychological Tests*. Washington, DC: APA.

The two quotations from Gardner come from

Gardner, E.F. (1978) Bias. *Measurement in Education*, 9 (3).

The quotations from Haugher can be found in

Haugher, R.L. (1978) The many definitions of test bias. *American Psychologist*, 33(7), pp. 671–679.

The quotation from Silverman can be found in

Silverman, B. (1979) Test bias and ability level testing. *Journal of School Psychology*, 17 (3), pp. 255–259.

Gulliksen's comments on biased criteria can be found in

Gulliksen, H. (1976) *When High Validity May Indicate a Faulty Criterion, (RM76–10)*. Princeton, NJ: ETS.

The quotation from Reschley can be found in

Reschley, D.J. (1978) *Non-biased Assessment*. Des Moines, Iowa: State of Iowa Department of Public Instruction, p. 33.

The quotation from Darlington can be found in

Darlington, R.B. (1971) Another look at cultural fairness. *Journal of Educational Measurement*, 8, pp. 71–82.

The idea that political considerations, not technical ones, will probably carry more weight comes from

Mercer, J.R. (1979) Test 'validity', 'bias', and 'fairness': an analysis from the perspective of the sociology of knowledge. *Interchange*, 9(1), pp. 1–16.

The last quotation is from

Ebel, R.L. (1963) The social consequences of educational testing. *Proceedings of the 1963 Invitational Conference on Testing Problems*. Princeton, NJ: ETS, pp. 142–143.

The test usually known as the TOSCA is

Reid, N., Jackson, P., Gilmore, A. and Crott, C. (1981) *Test of Scholastic Abilities*. Wellington: NZCER.

Further Reading

For a comprehensive overview, see:

Cleary, T.A., Humphreys, L., Kendrick, S.A. and Wesman, A. (1975). Educational use of tests with disadvantaged students. *American Psychologist*, 30(1), pp. 15–41.

For a summary of tester effects as an aspect of bias, see:

Graziano, W.G. and Varca, P.E. (1982). Race of examiner effects and the validity of intelligence tests. *Review of Educational Research*, 52(4), pp. 469–497.

Matters of content bias in 'intelligence' tests are addressed in:

Zoref, L. and Williams, P. (1980). A look at content bias in IQ tests. *Journal of Educational Measurement*, 17(4), pp. 313–322.

These articles offer comment and discussion on bias in prediction selection:

Linn, R.L. (1982). Admissions testing on trial. *American Psychologist*, 37(3), pp. 279–291.

McNemar, Q. (1975). On so-called test bias. *American Psychologist*, 30(8), pp. 848–851.

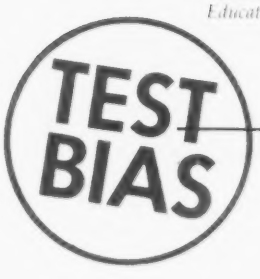
Four articles providing contrasting views of racial and socio-economic bias in 'intelligence' tests:

Gordon, R.A. and Rudert, E.E. (1979). Bad news concerning IQ tests. *Sociology of Education*, 52, pp. 174–190.

Guterman, S.S. (1979). IQ tests in research on social stratification: the cross-class validity of tests as measures of scholastic aptitude. *Sociology of Education*, 52, pp. 163–173.

Jensen, A.R. (1976). Test bias and construct validity. *Phi Delta Kappan*, 58(4), pp. 340–346.

Trotman, F.K. (1977). Race, IQ and the middle-class. *Journal of Educational Psychology*, 69(3), pp. 266–273.



Contents and Commentary
Cedric Croft

Assessment Issues and Measurement Concepts

Overview of Issues in School Assessment
Barry McGaw

Achievement Test Scores in Perspective
Bill Turnbull

Foundations of School Testing
Cedric Croft

Test Evaluation Sheet
Cedric Croft.

Measurement and Assessment Techniques

Assessing What They've Learned
Warwick Elley

Criterion Referenced Tests
Glenn Rowley and Colin McPherson

Investing in Item Banks
Neil Reid

Combining Scores
Alison Gilmore

Evaluating Writing
David Philips

Observation: The Basic Techniques
Bruce McMillan and Anne Meade

Reporting

One Extreme to the Other: A Report on Profile Reports
Graeme Withers

Assessment, Abilities and Culture

Non-Verbal Tests in Schools
Cedric Croft

Does Intelligence Equal Learning Ability?
Jo Jenkinson

Test Bias! Test Bias!
Neil Reid and Alison Gilmore